

nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE

**Antibiotics
target butterfly-
fold element
in a bacterial
riboswitch**

PAGES 650 & 672

DRUGGABLE NON-CODING RNA

CLIMATE FRONTLINE

THAT SINKING FEELING

*If rising tides don't kill
Kiribati, then thirst will*

PAGE 624

FOSSIL-FUEL EMISSIONS

MAKE MORE OF CARBON

*Can we turn CO₂ from
liability to valuable resource?*

PAGE 628

MICROBIAL ECOLOGY

ALL IN IT TOGETHER

*Global effort needed to study
Earth's microbiome*

PAGE 631

NATURE.COM/NATURE

29 October 2015 £10

Vol. 526, No. 7575



THIS WEEK



EDITORIALS

GHOST STORY The sinister tale of lost and abandoned fishing gear **p.610**

WORLD VIEW Beware the transfer of innocent people's DNA to crime scenes **p.611**

PLAGUE Ancient remains show that disease was around early **p.613**

Power struggle

The UK government's decision to subsidize a nuclear power station while cutting support for renewables is short-sighted.

The English poet William Blake once wrote that “Energy is an eternal delight”. But then poets have rarely been charged with keeping the lights on. Some 200 years later, energy — and how to produce and harness it — is on track to become the defining problem of a generation.

The latest dark satanic mill on the horizon in Blake's green and pleasant land is a shiny new nuclear power station — confirmed last week after years of plotting. It will be the first built in the United Kingdom this century, and is one of the most significant nuclear deals worldwide since the meltdown at the Fukushima Daiichi nuclear power plant in Japan in 2011.

As Chinese President Xi Jinping — in the United Kingdom for a state visit — was paraded around a series of picture-book British locations, it was announced that China was taking a 33.5% stake in the Hinkley Point C nuclear plant. French power company EDF will own the remaining 66.5%.

Supported by billions of pounds of Chinese investment, the plant should provide 3.2 gigawatts of power when it fires up as planned in 2025. But never mind the output: feel the cost.

The UK government has agreed a price with the investors of at least £89.50 (US\$137) in 2012 terms for every megawatt hour of power produced by the plant. This is roughly double the current market cost, but the government claims — with a little poetic licence of its own — that it “is competitive with other large energy sources such as gas and offshore wind”.

The Conservative government also noted that this would mean abandoning the policy of several previous administrations that there should be no public subsidy for new nuclear power.

In reality, it had little choice. (The previous administrations had no problems sticking to the line, because they had no new nuclear power to subsidize.) Nuclear power plants are among the single most expensive items that governments can buy, and as Britain has allowed its home-grown nuclear expertise to dwindle, so it has lost the chance to bury the exact cost in a tangle of public expenditure. Whether the money comes directly from the public as artificially high electricity bills, or indirectly through the public purse, British politicians last week admitted that the technology simply cannot pay for itself.

In doing so, they have been criticized — there are plenty of opponents who argue, with some justification, that this nuclear deal is a poor one for Britain. But in a way, the new-found candour about the costs does at last make the debate about energy a more honest one. Most power industries, to a greater or lesser extent, need a financial leg-up.

Hinkley Point C will cost at least £18 billion. And that does at least buy a reduction in greenhouse-gas emissions compared with one carbon-heavy alternative for baseload generation: coal. But the UK government has done little to stress that point — it is hard for it to do so while systematically cutting off other low-carbon forms of electricity at the knees.

Last year, renewable energy sources supplied a record 19% of UK electricity (the global figure is around 22%), but the same government that is committed to subsidizing Hinkley Point C has set out to slash subsidies for renewables. Trade associations say that thousands of jobs could go. Businesses face collapse as sections of the solar industry deal with 87% cuts in their subsidies, and onshore wind is under threat. Perversely, ministers justified the sacrifice of these growing clean-energy sources by pointing to increasing public support for them.

“Most power industries, to a greater or lesser extent, need a financial leg-up.”

The United Kingdom is far from alone in struggling to balance short-term financial prudence with long-term environmental protection. Earlier this month, the International Energy Agency reported that renewable energy accounts for 45% of new power capacity added globally in 2014. But it warned that the rate of growth for renewables was slowing because of policy problems and uncertainties — especially in Europe and Japan.

UK energy and climate-change minister Andrea Leadsom used the ugly word “trilemma” last week to describe the energy issue facing the United Kingdom: the need to reduce carbon emissions, to maintain supplies and to keep bills down.

There is another trilemma, popular among project managers: the need for any project to be good, fast and cheap. The classic response could have been written for energy policy: ‘pick two’. ■

Burst bubbles

Two medical-technology companies illustrate the ups and downs of innovation.

From time to time in most industries, the conventional approach is challenged by upstarts. Often backed by entrepreneurs and investors, these firms promise to use new technology to disrupt, overturn and revolutionize. Some succeed and some do not, and there are fields in which the challenge to newcomers is proving stiffer than others. One of these is health care, and events over the past week or so demonstrate both the difficulties and the opportunities.

Theranos and 23andMe are two medical-technology companies with their origins in Silicon Valley. Both have made headlines recently. Their stories may seem similar. But the differences offer an important lesson for would-be health disruptors: this industry can change, just not as quickly as entrepreneurs and their investors might hope, and only if those offering the change can also offer data to back up their claims.

Theranos in Palo Alto, California, promised to upend medicine with a device that can perform hundreds of diagnostic tests on just a few drops of blood. 23andMe, in Mountain View, California, sells genetic tests directly to consumers. Both are led by charismatic female founders: Elizabeth Holmes at Theranos and Anne Wojcicki at 23andMe. Both want to revolutionize the health-care industry and argue that patients should have access to their data. They have strong backing from Silicon Valley investors, and were hyped early on: a US\$9-billion valuation for Theranos, and lavish parties with media tycoons for 23andMe.

But both have seen their bubbles burst. On 16 October, *The Wall Street Journal* reported that the Theranos technology was not working as billed, and that the firm was using conventional machines to perform most of its tests. The company has disputed some of the article's claims. Holmes says that the company is now in a "pause period" because of scrutiny from US regulators.

23andMe's bubble burst in November 2013, when the US Food and Drug Administration (FDA) banned the inclusion of medically relevant results in the company's consumer genetic tests. On 21 October, however, 23andMe relaunched consumer genetic tests that give a limited amount of medical information, with FDA approval. The new tests offer information for 36 diseases about a customer's status as a 'carrier' of genetic glitches that could cause disease if passed down through their children.

Theranos could learn a lot from how 23andMe returned to the regulators' good graces. 23andMe has always been fairly open about its science; it publishes research papers in peer-reviewed journals and collaborates with scientists. Theranos, by contrast, has been tight-lipped about its data. Apart from detailed data for one herpesvirus test, approved by the FDA in July, the company has published only aggregate test performances on its website, not the primary data.

23andMe says that coming back from its early mistakes with

the FDA was an arduous process — requiring it to hire staff with expertise in health regulation and to compile detailed dossiers of data to prove that its tests work as advertised. The company previously had been slow to respond to the FDA's entreaties — and that tone-deafness seems to have been part of the reason that the agency eventually cracked down.

These experiences do not mean that health care cannot be disrupted. Indeed, 23andMe is the first company to gain FDA approval to sell a health-related genetic test to consumers without a doctor's order. That's a real change.

"Time and again, new health-care firms are forced to realize that it helps no one to be secretive with data."

Still, the new tests offer less information than before and at a higher price. With a few exceptions, carrier tests do not say anything about the health of the individual tested, and they are mainly for rare diseases — a far cry from the risk-prediction scores the company previously offered for cancer and Alzheimer's disease. And the new test package costs US\$199 compared with \$99 before the ban.

23andMe has also moved to make itself into a more conventional pharmaceutical firm. In March, it hired former Genentech executive Richard Scheller to lead a drug-development arm. If you can't beat them, join them.

Time and again, new health-care firms are forced to realize that it helps no one to be secretive with data. Even if it turns out that the Theranos technology does not work as well as advertised, the company would hardly be the first to find itself in that situation. Releasing more information earlier might have forced Theranos to confront shortcomings. Instead, it finds itself trying to recover from a regulatory and public-relations hole. This is not an insurmountable situation, as 23andMe knows. The challenge now is for Theranos to show us the data. ■

Ghost story

The problem of abandoned fishing gear and its effects on marine life deserve greater attention.

Here's a horror story for Halloween. Right now, in unlit waters across the world, fish, crabs and other marine life are being drawn into nets and traps by the dead and decomposing bodies of their comrades. There they will stick, struggle and tangle, until they, too, become unwitting bait and continue the circle of death. Old fishing nets, you see, never die. They just drift away.

The problem of ghost fishing, as it is known, goes largely unnoticed, but some of this dead gear catches and kills more sea life than it did when it was alive and in active use. Reliable data on the scale of the problem are scarce, but some estimates suggest that the nets can remove up to 30% of the landed catch of certain fish species.

It is said that we know more about the surface of the Moon than about the bed of the sea. Perhaps we are afraid of what we will find there if we look too hard: wrecks of gill nets, entangling nets, pelagic and demersal longlines, lobster and crab pots, seine nets, trawl-net fragments and the sinister-sounding fish aggregating devices — buoys or floats, tethered to concrete blocks, around which fish tend to congregate.

Some of this fishing gear is lost and some is abandoned in rough weather. Much is simply discarded by fishers with nowhere to stow it, who are fishing where they should not be, or who just want to avoid the expense and hassle of disposing of it properly. Most of this gear sinks to the bottom. It becomes a hazard, to boats and divers. And much of it continues to catch and kill, long after it has been forgotten.

Take the coastline of Louisiana, a US state that is home to its fair share of spooky tales. Each of the 1,800 or so professional crab fishers who work there loses about 250 traps every year. Each abandoned trap, a crude wire cage, is reckoned to catch and kill a blue crab (*Callinectes sapidus*) every two weeks. That is 12 million crabs a year, or 2 million kilograms of crab meat — about US\$4-million worth — along a single stretch of coastline (J. A. Anderson and A. B. Alford *Mar. Pollut. Bull.* **79**, 261–267; 2014). Ghost crab traps snare other creatures too: spotted sea trout, diamondback terrapins and river otters among them.

Although the world organizes regular conferences to address the threat of orbiting space junk, action on the danger of ghost fishing tends to be left to volunteers. Louisiana law allows a ten-day period each year when citizens can drag derelict fishing gear from the water. In two sessions — 2012 and 2013 — volunteers recovered a total of 3,607 ghost crab traps. More than 65% of them had caught something. The actions of such volunteers are admirable but they are not enough. The fishing industry and those who profit from it must take more responsibility.

Earlier this year, Eric Gilman, a fisheries scientist at Hawaii Pacific University in Honolulu, published a survey of international efforts to track and control ghost fishing (E. Gilman *Mar. Policy* **60**, 225–239; 2015). Of the 19 global and regional bodies (from the International Whaling Commission down to the South East Atlantic Fisheries Organization) that he identified as being in a position to intervene, just 4 had an explicit mandate to monitor and reduce the problem. Almost half did not even collect data on lost gear. The 12 organizations that have introduced measures to help prevent and reduce ghost fishing have not used all the options available to them.

All ghost stories are more chilling in the dark. The problem of abandoned, lost and discarded fishing gear deserves more attention and more action. For unlike many gruesome stories you will hear this weekend, this one is true. ■

➔ **NATURE.COM**
To comment online,
click on Editorials at:
go.nature.com/xhunqy

ERIN NELIS



Forensic DNA evidence is not infallible

As DNA analysis techniques become more sensitive, we must be careful to reassess the probabilities of error, argues Cynthia M. Cale.

Earlier this month, the Texas Forensic Science Commission raised concerns about the accuracy of the statistical interpretation of DNA evidence, and it is now checking whether convictions going back more than a decade are safe.

Despite how it is often portrayed, in the media and in courts, the forensic science of DNA is far from infallible. Particularly concerning is that police and prosecutors now frequently talk of 'touch DNA' — genetic profiles of suspects and offenders that have been generated in a laboratory from just a handful of skin cells left behind in a fingerprint.

Research done by me and others at the University of Indianapolis in Indiana has highlighted how unreliable this kind of evidence can be. We have found that it is relatively straightforward for an innocent person's DNA to be inadvertently transferred to surfaces that he or she has never come into contact with. This could place people at crime scenes that they had never visited or link them to weapons they had never handled.

Such transfer could also dilute the statistics generated from DNA evidence, and thereby render strong genetic evidence almost insignificant. (This issue is the focus of the Texas investigation.)

We urgently need to review how DNA evidence is assessed, viewed and described. Everyone in the medico-legal community — forensic scientists and technicians, DNA analysts, potential jurors, judges and lawyers for both the prosecution and defence — must know and understand the potential for mistakes.

The term 'touch DNA' conveys to a courtroom that biological material found on an object is the result of direct contact. In fact, forensic scientists have no way of knowing whether the DNA was left behind through such primary, direct transfer. It could also have been deposited by secondary transfer, through an intermediary. (If I shake your hand then I could pass some of your skin cells onto something that I touch next.)

Contamination from secondary DNA transfer was raised as a possible problem in *Nature* in 1997 (R. A. H. van Oorschot and M. K. Jones *Nature* **387**, 767; 1997). It is known to happen, but has largely been dismissed by legal experts as being rare outside the conditions of a laboratory. Experiments done in real-world conditions seemed to support this, and concluded that secondary DNA transfer would have little impact on interpretation of the genetic profile.

It is important to recognize that DNA amplification kits have become much more sensitive than they were in the past.

As a result, the types of samples being analysed have expanded. Investigators no longer need to identify and request analysis of body fluids such as blood, semen and saliva. They can swab

surfaces for otherwise invisible cells left behind, on the handle of a weapon or on a windowsill, perhaps, and ask labs to generate a DNA profile from them. The new kits can generate a full genetic profile of a suspect from as little as 100 picograms (trillionths of a gram) of DNA.

These subtleties are not usually explained in court. Instead, a jury is told that there is a one-in-a-quadrillion chance that the evidence retrieved from the crime scene did not come from a defendant. Naturally, the jurors assume that the defendant must have been there.

Given the power of modern forensic techniques to pull a DNA profile from a smudge of cells, secondary DNA transfer is no longer a purely theoretical risk. In California in 2013, a man called Lukis Anderson was arrested, held for four months and charged with murder after his DNA was found under the fingernails of a homicide victim.

Anderson had never met the victim and was severely intoxicated and in hospital when the man was killed. The same paramedics who took Anderson to hospital responded to the murder. Most likely, the paramedics were covered in Anderson's DNA, which they then inadvertently transferred. The charges were dropped.

Experiments in our labs, under the supervision of forensic anthropologist Krista Latham, show how easily DNA can be transferred to an object.

We asked pairs of people to shake hands for two minutes and then each individual handled a separate knife. In 85% of cases, the DNA of the other person was transferred to the knife and profiled. In one-fifth of the samples, the DNA analysis identified this other person as the main

or only contributor of DNA to the 'weapon' (C. M. Cale *et al. J. Forensic Sci.* <http://doi.org/8j2>; 2015).

How significant is the result of a single study? Other analyses have shown that DNA transfer can be unpredictable and can depend on environmental conditions. We need more research on when and how secondary transfer can occur.

At the very least, the results highlight again that samples at crime scenes must be gathered with great care. DNA can persist on latex gloves, so they must be changed — or bleached — before and after handling evidence.

Even apparently rigorous evidence such as DNA profiles can be interpreted in multiple ways, some of which will be incorrect. As the technology to generate these profiles continues to accelerate, so must our efforts to sift out possible mistakes. ■

Cynthia M. Cale is a human-biology graduate student at the University of Indianapolis, Indiana, and lead forensic DNA analyst III at Strand Diagnostics.
e-mail: ccale@strandlabs.com

**WE NEED
MORE
RESEARCH
ON WHEN AND HOW
SECONDARY
TRANSFER
CAN OCCUR.**

➔ **NATURE.COM**
Discuss this article
online at:
go.nature.com/wojxtm

RESEARCH HIGHLIGHTS

Selections from the
scientific literature

NEUROSCIENCE

Alzheimer's clue from spatial test

Young adults who are at increased risk of developing Alzheimer's disease show abnormal function in a part of the brain involved in spatial navigation.

Nikolai Axmacher at Ruhr University Bochum in Germany and his colleagues used neuroimaging to measure the functioning of the 'grid-cell' system in the entorhinal cortex as participants with or without the *APOE-ε4* risk gene navigated a virtual arena.

The 38 high-risk individuals showed reduced grid-cell functioning compared with the 37 controls and tended to avoid the centre of the arena. Activity in the hippocampal region was higher when the grid-cell system was impaired, as if to compensate for the deficit. Reduced grid-cell functioning may prove useful as an early biomarker of Alzheimer's, the authors suggest.

Science 350, 430–433 (2015)

ASTRONOMY

Red-giant rogue in Andromeda

Astronomers have spotted a giant 'runaway star' speeding through the Andromeda galaxy; the first of its kind seen outside the Milky Way.

Whereas most stars flow together around the centre of their galaxy, some, known as runaways, travel at different directions and speeds to their peers, and may even escape their galaxy entirely. Spotting red-supergiant runaways is rare — stars in this late stage of life left their birthplace long ago, making their unusual speeds harder to spot than for younger runaways.

Kate Evans and Philip



ZOOLOGY

Bright light as sex signal

Brighter female glow-worms lay more eggs than their dim rivals and are more attractive to potential nocturnal mates.

Juhani Hopkins at the University of Oulu in Finland and his colleagues allowed 26 female glow-worms (*Lampyris noctiluca*; pictured) to mate in the lab. The glowing lanterns of the insects varied in size from 7 square millimetres to 19 square millimetres — larger lanterns produce a brighter glow. Each glow-worm laid between 25 and 195 eggs, with those perceived by the researchers to be brightest laying the most. Male glow-worms presented with fake females also preferred those with brighter lights.

The lanterns of female glow-worms may provide clues about fitness to males, who are unable to assess size — also an indicator of fecundity — in the dark.

Biol. Lett. <http://dx.doi.org/10.1098/rsbl.2015.0599> (2015)

Massey at the Lowell Observatory in Flagstaff, Arizona, studied a red supergiant in the Andromeda galaxy known as J004330.06+405258.4 and calculated that it is travelling 400–450 kilometres

per second faster than its neighbours. The star is the first massive runaway to be spotted outside our own galaxy and the fastest anywhere for its size, say the authors.

Astron. J. 150, 149 (2015)

MATERIALS

Iron skin senses the softest touch

An iron-based artificial skin can sense the lightest touch.

Ahmed Alfadhel and Jürgen Kosel at the King Abdullah University of Science and Technology in Thuwal, Saudi Arabia, made a tactile sensor by embedding iron nanowires in hair-like structures called cilia, made of a polymer called polydimethylsiloxane, on a magnetic sensor. When the magnetized cilia are bent by touch, they trigger the sensor.

The skin can flex and, depending on the size of the cilia, can be sensitive enough to measure a person's wrist pulse. Because the cilia use permanent magnets rather than electromagnets, the device does not require much power and — unlike most other tactile sensors — it can work underwater and measure liquid flow, the authors say.

Adv. Mater. <http://doi.org/f3jp2n> (2015)

GENOMICS

Gene regulation predates animals

The oldest ancestor of animal life used the same tricks that modern humans do to turn genes on and off.

Alex de Mendoza at the Institute of Evolutionary Biology in Barcelona, Spain, and his colleagues studied gene regulation in the fungus-like single-celled organism *Creolimax fragrantissima*, which branched onto a separate evolutionary path before the evolution of multicellular organisms.

To produce different cell types, multicellular organisms use three main gene-regulation processes: transcription factors, alternative splicing and

ANDY SANDS/NATURE PICTURE LIBRARY

non-coding RNAs. The authors found that *C. fragrantissima* uses the same processes to switch between life stages, meaning that these regulatory elements were likely to have been used by the last universal common ancestor of all animals, the authors say.

eLife <http://doi.org/8kh> (2015)

DISEASE

Plague is an ancient pathogen

Plague was plaguing humanity thousands of years earlier than previously thought, but in a less transmissible form.

Yersinia pestis bacteria, which are thought to have been behind the Black Death that killed millions in the fourteenth century, have previously been found in burial sites dating back 1,500 years.

But Eske Willerslev at the Natural History Museum of Denmark in Copenhagen and his colleagues looked even further back. They analysed DNA obtained from the teeth of 101 humans (pictured) who died in Europe and Asia between 2,800 and 5,000 years ago and found *Y. pestis* DNA in seven individuals. Analysis of the DNA showed that a strain similar to the Black Death strains was widespread in the Bronze Age, but only the more recent strains had a gene called *ymt*, which helps *Y. pestis* to colonize the guts of fleas. Without fleas to aid transmission, plague spreads less efficiently.

Cell <http://dx.doi.org/10.1016/j.cell.2015.10.009> (2015)

JOE RAEDLE/GETTY IMAGES

RASMUSSEN ET AL./CELL 2015



STEM CELLS

Molecular menu creates neurons

Astrocyte cells in the brain can be reprogrammed into neurons using a precise sequence of molecules. The technique may one day be useful in brain repair.

Similar cells have previously been reprogrammed into neurons using viruses, but Gong Chen and Gang-Yi Wu at Pennsylvania State University in University Park and their colleagues now show that the transformation can be done with small molecules.

They treated human astrocytes with nine different molecules in sequence, converting them into neurons that survived for more than five months in culture and more than one month after transplantation into a mouse brain. The method works for human brain astrocytes but not for human spinal astrocytes or mouse astrocytes, suggesting that different sets of molecules may be needed for different astrocytes or to obtain different neuronal subtypes, the authors report.

Cell Stem Cell <http://doi.org/8m5> (2015)

LAB TOOLS

Superconducting sensors warm up

An extremely sensitive, superconductor-based magnetic sensor can work at around 77 kelvin, a temperature achievable with liquid nitrogen rather than the expensive liquid helium required by typical existing devices, which operate at just above absolute zero.

Superconducting quantum-interference devices (SQUIDs) can sense individual quanta of magnetic flux by measuring voltage induced in a loop of superconducting material. Boris Chesca of Loughborough University, UK, and his team connected hundreds of loops in series to boost the signal. The authors

SOCIAL SELECTION

Popular topics on social media

Backlash over journals blacklisting

Researchers on social media are split over the decision of academic librarian Jeffrey Beall to add the Frontiers journals to his 'blacklist' of "questionable publishers". Beall, at the University of Colorado Denver, announced the move in a tweet, saying that it followed "wide disapproval from scientists". His website Scholarly Open Access maintains a list of journals that may be "predatory publishers" — a term Beall coined for publications that charge scientists fees to publish research but that do not offer services such as peer review, or that make misleading claims on impact factors or indexing. Critics of Beall's blacklisting of Frontiers maintain that the open-access publisher is reputable and does offer proper peer review. Daniël Lakens, an experimental psychologist at the Eindhoven University of Technology in the Netherlands and an associate editor at *Frontiers in Psychology: Cognition*,

↪ **NATURE.COM**
For more on
popular papers:
go.nature.com/ch66au

tweeted: "Frontiers being added to Beall's list reveals the big weakness of Beall's list: It's not based on solid data, but on Beall's intuition." Beall told *Nature* that he stands by his decision.



say that their SQUID design is as sensitive as many devices already in use and is ready for production. The higher operating temperature makes it ideal for applications such as portable magnetic resonance imaging machines, they say.
Appl. Phys. Lett. 107, 162602 (2015)

ATMOSPHERIC SCIENCE

Arctic snow is not becoming dirtier

Dust and soot might not be behind the observed darkening of the Greenland ice sheet (pictured).

Tiny particles of dirt absorb sunlight that would be reflected into space by ice — contributing to local warming. Satellite measurements suggest that the amount of sunlight reflected by Greenland's icy

surface has been decreasing since 2001.

But surveys of the snow in northwest Greenland conducted in 2013 and 2014 by Chris Polashenski at the US Army Cold Regions Research and Engineering Laboratory in Fort Wainwright, Alaska, and his colleagues, found that the concentration of dark-coloured particles was much the same as in previous decades. Rather than dirtier ice, the declining reflectivity seen in satellite measurements could be due to a degrading sensor on NASA's Earth-observing satellite Terra.
Geophys. Res. Lett. <http://doi.org/8kg> (2015)

↪ **NATURE.COM**
For the latest research published by *Nature* visit:
www.nature.com/latestresearch

SEVEN DAYS

The news in brief

EVENTS

Afghan quake

At least 300 people have been killed, with the number expected to rise, after a magnitude-7.5 earthquake hit northeast Afghanistan on 26 October. The strong quake, which according to the US Geological Survey struck 76 kilometres south of the city of Fayzabad at a depth of around 210 kilometres, caused landslides and avalanches in the remote Hindu Kush region and sent tremors that were felt in neighbouring Pakistan and as far away as Delhi in India. Earthquakes in the tectonically active region result from the collision of the Indian subcontinent with central Asia.

Science subpoenaed

The chairman of a US congressional committee has issued a subpoena requesting to see all documents and communications related to a June study by federal scientists showing that global warming has continued unabated over the past 15 years. Lamar Smith (Republican, Texas), chair of the Committee on Science, Space, and Technology, issued the subpoena on 13 October after the National Oceanic and Atmospheric Administration refused to provide internal communications. The subpoena came to light publicly on 23 October after Democrat Bernice Johnson from Texas objected to the “illegitimate harassment of our Nation’s research scientists”.

Meltdown payout

Japan’s Ministry of Health, Labour and Welfare has said that it will pay industrial-accident compensation to a member of the clean-up crew who worked on the damaged Fukushima Daiichi nuclear power plant and who has

recently developed cancer. According to the Japanese media, the ministry said on 20 October that the causal link between the worker’s exposure and his leukaemia was “unclear”. The government could be exposed to many more claims. Thousands of workers took part in the clean-up operations after three reactors went into meltdown in March 2011, following an earthquake and a tsunami.

POLICY

Bonn climate talks

A draft text for a global greenhouse-gas-reduction agreement was drawn up at talks in Bonn, Germany, between 19 and 23 October. During the talks, Sri Lanka and the United Arab Emirates

became the 155th and 156th countries to submit pledges for mandatory domestic action, which will be at the core of any United Nations climate agreement made in Paris this December. Meanwhile, in a briefing released on 21 October, the International Energy Agency said that US\$13.5 trillion need to be invested globally in energy-efficiency and low-carbon technologies over the next 15 years to compensate for the planned reduction in fossil-fuel use.

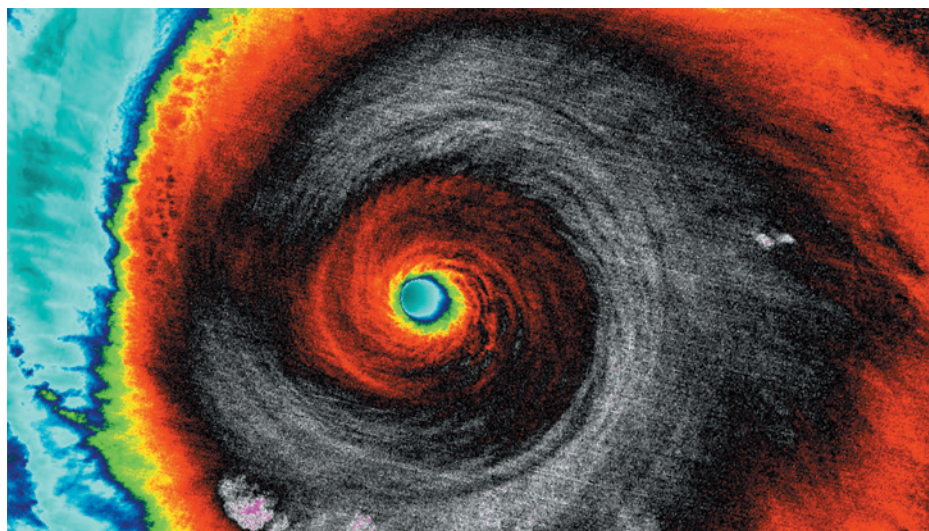
Marine reserve

The Congress of Palau approved plans to create a 500,000-square-kilometre marine reserve around the Pacific island nation on 22 October. The

announcement came in the same week as an analysis showed that the amount of the world’s oceans that is ‘strongly protected’, with some fishing allowed, or ‘fully protected’ has increased from 0.1% to 1.6% in the past decade (J. Lubchenco and K. Grorud-Colvert *Science* **350**, 382–383; 2015). The area that is afforded some level of protection is 3.5%, still well short of the internationally agreed target of 10% by 2020.

Climate regulation

Pakistani Prime Minister Nawaz Sharif endorsed proposals to regulate hydrofluorocarbons (HFCs) under the Montreal Protocol on Substances that Deplete the Ozone Layer. Commonly used as refrigerants, HFCs were created as alternatives to



UW/CIMSS/WILLIAM STRAKA III

Hurricane Patricia smashes records

Hurricane Patricia broke a slew of meteorological records just before it pounded the west coast of Mexico on 23 October (pictured in an infrared satellite image). The category-5 storm intensified rapidly before making landfall, with wind speeds of almost 324 kilometres per hour and a central pressure of 87.9 kilopascals, making

it the strongest hurricane ever recorded in the Western Hemisphere. Heavy rains raked the coast between the cities of Puerto Vallarta and Manzanillo, causing landslides, but the devastation was less than feared because the most intense winds and storm surge struck in relatively lightly populated areas.

chlorofluorocarbons, which damage the ozone layer — but HFCs are still powerful greenhouse gases. Sharif declared his support in a joint statement with US President Barack Obama on 22 October after a visit to the White House. At a meeting in Dubai on 1–5 November, parties to the Montreal Protocol will consider amendments intended to phase out HFCs.

PEOPLE

Science top job

Australia's new prime minister, Malcolm Turnbull, appointed engineer Alan Finkel as chief scientist on 27 October. Finkel, who co-founded the popular-science magazine *Cosmos* and is currently chancellor of Monash University, is a vocal advocate of nuclear power as a carbon-neutral energy source, and of the need for policies to combat climate change. Media reports suggest that Finkel's appointment may signal a policy shift in a country in which nuclear energy is banned and whose previous prime minister was a notorious climate-change sceptic. Finkel said in a press conference that he wanted to see the end of the use of coal, oil and gas in the country, with renewables and nuclear power being explored to achieve zero-emissions energy.



Lisa Jardine dies

Historians of science, politicians and researchers paid tribute on 26 October after it was announced that historian Lisa Jardine (pictured) had died. Jardine's hugely varied career included work on eighteenth-century scientist Robert Hooke and the scientific revolution. She chaired Britain's Human Fertilisation and Embryology Authority between 2008 and 2014, during a period when the agency led the way in regulating mitochondrial replacement and other cutting-edge scientific advances. Mitochondrial-replacement therapy was made legal in Britain in February 2015.

RESEARCH

Mars lander

Europe's first Mars rover, ExoMars, looks set to land on a vast, clay-rich plain called Oxia Planum, the

European Space Agency (ESA) announced on 21 October. ExoMars, a joint endeavour between ESA and Russia's space agency Roscosmos, will land in 2019 and is designed to look for evidence of life. The craft's instruments will include a drill capable of burrowing 2 metres into the Martian surface. Oxia Planum was picked from a shortlist of four potential landing sites (see *Nature* 508, 19–20; 2014); the agencies will confirm the choice six months before the mission's 2018 launch.

FACILITIES

Telescope closed

The United Kingdom Infrared Telescope (UKIRT) atop Mauna Kea in Hawaii will be permanently shuttered, making it the third closed to accommodate the next-generation Thirty Meter Telescope (TMT). The University of Hawaii took over the UKIRT — the most scientifically productive ground-based telescope worldwide — from the UK Science and Technology Facilities Council last year. The decision to close stems from a long-term decommissioning plan, which identifies the UKIRT as a site to be restored to its natural state. Hawaii's governor has called for closure of one-quarter of Mauna Kea's 13

COMING UP

1–4 NOVEMBER

Researchers hoping to catch up on the latest in cancer genomics can do so at a European Molecular Biology Laboratory conference in Heidelberg, Germany.

go.nature.com/p5uqwp

1–4 NOVEMBER

The Geological Society of America holds its annual meeting in Baltimore, Maryland.

go.nature.com/ks35w9

4–7 NOVEMBER

Budapest hosts the 2015 World Science Forum, themed 'the enabling power of science'.

www.sciforum.hu

telescopes by the time the TMT becomes operational in the 2020s (see *Nature* 526, 24–28; 2015).

BUSINESS

23andMe revamp

Genetic-testing company 23andMe, based in Mountain View, California, launched a revamped consumer test on 21 October. The product offers information about a customer's genetic-carrier status for 36 diseases, marking the first time that 23andMe has been allowed to provide medically relevant results since the US Food and Drug Administration banned the inclusion of detailed predictive information in tests in November 2013. Previously, the product tested for 240 health conditions. With a few exceptions, carrier-status tests do not say anything about the health of the individual and are mostly for rare diseases. See go.nature.com/5vicei and page 609 for more.

➔ NATURE.COM

For daily news updates see:

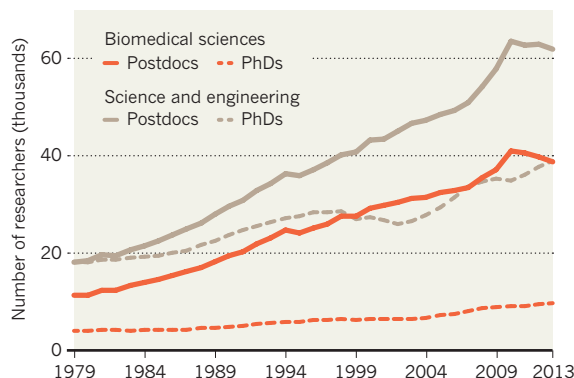
www.nature.com/news

TREND WATCH

The number of US biomedical postdocs fell 5.5% between 2010 and 2013, to just over 38,000, with losses getting bigger each year, according to an analysis (H. H. Garrison *et al.* *FASEB J.* <http://doi.org/8m3>; 2015). The main reason for the drop seems to be fewer positions for postdocs, but the time that individuals spend as postdocs is also falling. From 1979 to 2010, the number of US postdocs in the biomedical sciences rose steadily, from just over 10,000 to more than 40,000. See go.nature.com/fbkl5y for more.

POSTDOCS IN DECLINE?

The number of researchers in US postdoctoral positions is starting to fall, particularly in the biomedical sciences — although the number of new PhDs is still rising.



NEWS IN FOCUS

ANTARCTIC SUPER-DRILLS A faster way to bore into Earth's history **p.618**

EPIDEMIOLOGY Demise of UK baby study prompts questions **p.620**

VIRAL HELPERS Cancer-killing herpesvirus on brink of approval **p.622**



CHEMISTRY Scientists seek ways to turn CO₂ emissions into useful products **p.628**

KAREL PRINSLOO/AP/PRESS ASSOCIATION IMAGES



Trials of the world's first vaccine against malaria in sub-Saharan Africa have shown that it offers only modest protection.

PARASITIC DISEASE

Vaccine gets cautious boost

Malaria immunization endorsed for small-scale use in Africa.

BY EWEN CALLAWAY & AMY MAXMEN

The world's first vaccine against malaria should be rolled out in limited pilot demonstrations in Africa, an advisory group to the World Health Organization (WHO) in Geneva said on 23 October.

The decision — which the WHO's director-general is expected to formally endorse in November — follows 28 years of development by the London-based drug firm GlaxoSmithKline (GSK) and other backers including the Bill & Melinda Gates Foundation in Seattle, Washington; together they have spent US\$565 million on the drug.

The demonstrations, involving up to 1 million children, are needed because the vaccine is ineffective against malaria unless

children receive four doses spread out over 18 months, and even then it offers only modest protection.

"If this vaccine is not effective and we use it widely, we have spent a ton of money where it could be better placed," said Jon Abramson, a paediatric infectious-disease specialist at Wake Forest School of Medicine in Winston-Salem, North Carolina, and chair of the WHO Strategic Advisory Group of Experts (SAGE) on Immunization, in a press briefing.

The imperfections of the vaccine, called RTS,S and sold as Mosquirix, are well known: trials in more than 15,000 children, who were followed for up to four years in seven countries in sub-Saharan Africa, found that a series of four shots reduced the number of malaria cases by only 36% in young children, and by 26% in

infants¹. Still, even such a modest effect could be significant because malaria kills around half a million people annually, mostly children, and all other candidate vaccines are in much earlier stages of development.

The advisory group recommended a series of 3–5 pilot demonstrations in areas with a medium to high incidence of malaria, involving up to 1 million children in total. These will reveal whether parents bring their children back for all four doses of the vaccine, said Abramson. The pilots will also investigate safety issues associated with the vaccine, such as the potential to develop meningitis.

The demonstrations could start in 2016 and are expected to last 3–5 years. Seth Berkley, head of Gavi, the Vaccine Alliance, in Geneva, Switzerland, says that his organization will ►

► soon decide whether it will help to pay for the pilot demonstrations. “It certainly is possible that the board will say yes to this, but there’s no guarantee,” he says. As data roll in, SAGE will review its position: a final decision on whether to recommend deploying the vaccine more widely could come during this period.

Even if that happens, it is not clear what the uptake would be. Although African malaria-control officials welcome RTS,S, they say that they would need more funding to deploy the vaccine. Budgets for malaria prevention and treatment using measures such as insecticide-treated bed nets and artemisinin-combination therapies are already stretched thin.

GSK says that it will charge \$1–10 per shot, covering the company’s manufacturing costs and a return of 5%, to be reinvested in new vaccines for malaria or other diseases that are common in the developing world. But funding will also be needed to deliver the vaccine to children and for programmes to disseminate information. Parents must understand that their children can still get malaria even with the vaccine, says James Tibenderana, development director at the Malaria Consortium in Uganda.

POOR MATCH

SAGE’s decision to pilot the vaccine follows the publication of a study on 21 October, which revealed² that the vaccine’s poor performance in clinical trials is in part because it mimics a strain of the malaria parasite *Plasmodium falciparum* that is not commonly found in Africa.

The vaccine is composed partly of a fragment of circumsporozoite (CS) protein, which is found on the surface of the parasite. People who have been given RTS,S build up some immunity to malaria. But different parasites have slightly different CS proteins — and the study showed that fewer than 10% of parasites infecting some 5,000 children in the trials matched the CS protein in RTS,S. If the vaccine could be re-engineered to include bits of several surface proteins, it would be more effective, says Dyann Wirth, an infectious-disease researcher at the Harvard T. H. Chan School of Public Health in Boston, Massachusetts, who led the 21 October study.

That redevelopment could take years, although some researchers have been discussing the possibility, according to David Kaslow, who oversees the vaccine’s development at the non-profit health organization PATH. “It’s not trivial to tweak the vaccine to match the prevalent strains in an area,” he says, “but it’s not impossible.”

In the meantime, the advice to run demonstrations of RTS,S sends the right message, says Adrian Hill, a vaccinologist at the University of Oxford, UK. “What the field needs is other players to come forward and accelerate their more modern vaccine candidates.” ■

1. RTS,S Clinical Trials Partnership. *Lancet* **386**, 31–45 (2015).

2. Neafsey, D. E. et al. *N. Engl. J. Med.* <http://dx.doi.org/10.1056/NEJMoa1505819> (2015).



Ice cores allow scientists to analyse past precipitation and temperature changes.

CLIMATE SCIENCE

Super-fast drills hunt for oldest ice

Researchers will test machines that can penetrate kilometres in days rather than years.

BY ALEXANDRA WITZE

Drilling through ice sheets is a tedious task. It takes years of fieldwork to retrieve long ice cores that keep a continuous record of the climate stretching back hundreds of thousands of years.

Now there is a faster way to bore deep into Earth’s history. Anxious to get to ice as old as 1.5 million years, nearly double the age of the oldest existing core, climate researchers have developed a new generation of ‘rapid-access’ ice drills. Some of these rigs will face their first major tests during the Antarctic field season that begins this month.

These speedy tools take roughly a week, rather than years, to penetrate several kilometres of ice. They blitz through the top-most layers of ice to reach the ancient freeze beneath, where tiny bubbles of trapped air serve as a time capsule of environments long vanished.

One of the biggest and most ambitious machines, a US project known as the Rapid Access Ice Drill (RAID), is being shipped

in November from its construction site in Salt Lake City, Utah, to McMurdo Station in Antarctica (see ‘Climate clues’). The British Antarctic Survey will test a much smaller drill, also named RAID (for Rapid Access Isotope Drill) in December at the Sky Blu station on the Antarctic peninsula. French and Swiss research teams are developing their own fast drill designs.

The drills sacrifice detail for speed, however. They chip up or melt the ice as they go, so extracting an intact core is impossible. But these fast drills will be able to do quick surveys of places where researchers might return in future field seasons to extract a full ice core at a more leisurely pace. The US\$10.5-million US RAID drill, for instance, is designed to plough through more than 3 kilometres of ice in about a week. That speed would allow it to hop around Antarctica and drill several exploratory holes per season — instead of one hole over several seasons.

Even so, finding the planet’s most ancient ice will not be easy. “We’re looking for a very fortuitous set of circumstances that allow for

NICK COBBING/SHUTTERSTOCK/REX

CORRECTION

The News story 'Vaccine gets cautious boost' (*Nature* **526**, 617–618; 2015) incorrectly stated that David Kaslow was involved in the early development of RTS,S.

► soon decide whether it will help to pay for the pilot demonstrations. “It certainly is possible that the board will say yes to this, but there’s no guarantee,” he says. As data roll in, SAGE will review its position: a final decision on whether to recommend deploying the vaccine more widely could come during this period.

Even if that happens, it is not clear what the uptake would be. Although African malaria-control officials welcome RTS,S, they say that they would need more funding to deploy the vaccine. Budgets for malaria prevention and treatment using measures such as insecticide-treated bed nets and artemisinin-combination therapies are already stretched thin.

GSK says that it will charge \$1–10 per shot, covering the company’s manufacturing costs and a return of 5%, to be reinvested in new vaccines for malaria or other diseases that are common in the developing world. But funding will also be needed to deliver the vaccine to children and for programmes to disseminate information. Parents must understand that their children can still get malaria even with the vaccine, says James Tibenderana, development director at the Malaria Consortium in Uganda.

POOR MATCH

SAGE’s decision to pilot the vaccine follows the publication of a study on 21 October, which revealed² that the vaccine’s poor performance in clinical trials is in part because it mimics a strain of the malaria parasite *Plasmodium falciparum* that is not commonly found in Africa.

The vaccine is composed partly of a fragment of circumsporozoite (CS) protein, which is found on the surface of the parasite. People who have been given RTS,S build up some immunity to malaria. But different parasites have slightly different CS proteins — and the study showed that fewer than 10% of parasites infecting some 5,000 children in the trials matched the CS protein in RTS,S. If the vaccine could be re-engineered to include bits of several surface proteins, it would be more effective, says Dyann Wirth, an infectious-disease researcher at the Harvard T. H. Chan School of Public Health in Boston, Massachusetts, who led the 21 October study.

That redevelopment could take years, although some researchers have been discussing the possibility, according to David Kaslow, who oversees the vaccine’s development at the non-profit health organization PATH. “It’s not trivial to tweak the vaccine to match the prevalent strains in an area,” he says, “but it’s not impossible.”

In the meantime, the advice to run demonstrations of RTS,S sends the right message, says Adrian Hill, a vaccinologist at the University of Oxford, UK. “What the field needs is other players to come forward and accelerate their more modern vaccine candidates.” ■

1. RTS,S Clinical Trials Partnership. *Lancet* **386**, 31–45 (2015).

2. Neafsey, D. E. et al. *N. Engl. J. Med.* <http://dx.doi.org/10.1056/NEJMoa1505819> (2015).



Ice cores allow scientists to analyse past precipitation and temperature changes.

CLIMATE SCIENCE

Super-fast drills hunt for oldest ice

Researchers will test machines that can penetrate kilometres in days rather than years.

BY ALEXANDRA WITZE

Drilling through ice sheets is a tedious task. It takes years of fieldwork to retrieve long ice cores that keep a continuous record of the climate stretching back hundreds of thousands of years.

Now there is a faster way to bore deep into Earth’s history. Anxious to get to ice as old as 1.5 million years, nearly double the age of the oldest existing core, climate researchers have developed a new generation of ‘rapid-access’ ice drills. Some of these rigs will face their first major tests during the Antarctic field season that begins this month.

These speedy tools take roughly a week, rather than years, to penetrate several kilometres of ice. They blitz through the top-most layers of ice to reach the ancient freeze beneath, where tiny bubbles of trapped air serve as a time capsule of environments long vanished.

One of the biggest and most ambitious machines, a US project known as the Rapid Access Ice Drill (RAID), is being shipped

in November from its construction site in Salt Lake City, Utah, to McMurdo Station in Antarctica (see ‘Climate clues’). The British Antarctic Survey will test a much smaller drill, also named RAID (for Rapid Access Isotope Drill) in December at the Sky Blu station on the Antarctic peninsula. French and Swiss research teams are developing their own fast drill designs.

The drills sacrifice detail for speed, however. They chip up or melt the ice as they go, so extracting an intact core is impossible. But these fast drills will be able to do quick surveys of places where researchers might return in future field seasons to extract a full ice core at a more leisurely pace. The US\$10.5-million US RAID drill, for instance, is designed to plough through more than 3 kilometres of ice in about a week. That speed would allow it to hop around Antarctica and drill several exploratory holes per season — instead of one hole over several seasons.

Even so, finding the planet’s most ancient ice will not be easy. “We’re looking for a very fortuitous set of circumstances that allow for

NICK COBBING/SHUTTERSTOCK/REX

SOURCE: H. FISCHER ET AL. *CLIM. PAST* 9, 2489–2505 (2013)

DEEP FREEZE

Antarctica may harbour ice up to 1.5 million years old that could help scientists to understand Earth's past climate.



SOURCE: US RAID PROJECT

the preservation of very old ice,” says Jeffrey Severinghaus, a palaeoclimatologist at the Scripps Institution of Oceanography in La Jolla, California. Ideally, scientists would discover a thick sequence of ice layers, undisturbed by flowing glaciers, that has not been heated too much by the rock below. Possible locations include several of the high-elevation Antarctic ice domes, such as Dome A near to which China's Kunlun research station sits, or Dome C, where European researchers took five years to extract a core that reached into 800,000-year-old ice layers (see ‘Deep freeze’).

Now researchers want to push even further, to ice that is at least 1.2 million years old. That would provide data on an important shift in Earth's climate, when the planet's glacial cycles changed from being dominated by a 100,000-year pattern to 41,000-year cycles (H. Fischer *et al. Clim. Past* 9, 2489–2505; 2013).

Knowing what controlled that switch — and whether rising carbon dioxide levels played a part, along with factors such as changes in Earth's rotational tilt — would help scientists to better understand how ice sheets will behave as the world warms. “If we don't understand this we really don't understand the climate that we have today,” says Severinghaus.

The US and UK drills take different approaches to reach deep into Antarctica's past. Once the US RAID reaches the bottom of the ice sheet, it could drill up to 50 metres into the underlying rock. Analysing that rock could reveal when it was last exposed to cosmic rays — which, in turn, reveals the age of the overlying section of the Antarctic ice sheet. The first full-scale field trial for RAID is scheduled for 2016–17.

The British RAID is a much more modest project that costs less than £500,000 (US\$770,000) and uses a modified conventional ice-core drill. It will be able to penetrate only about 600 metres into the ice sheet, to ice that is 30,000 to 40,000 years old — but unlike the US RAID, it does not require drilling fluid, the weight of which adds considerably to the cost of moving a drill around. “You can't dry-drill deeper than that,” says Julius Rix, an engineer who is leading the drill's development. “But there are plenty of places of interest.”

A third drill, on a similar scale to the US machine, is the SUBGLACIOR probe being developed at Joseph Fourier University in Grenoble, France. This €3.2-million (US\$5.3-million) project aims to melt rather than chip its way through the ice sheet, measuring chemical isotopes of the melted water as it goes, to calculate the age of the ice. The drill would be able to penetrate several kilometres deep; full testing is slated for 2016–17 at the Concordia research station in Antarctica, says Olivier Alemany, a polar engineer at Joseph Fourier University.

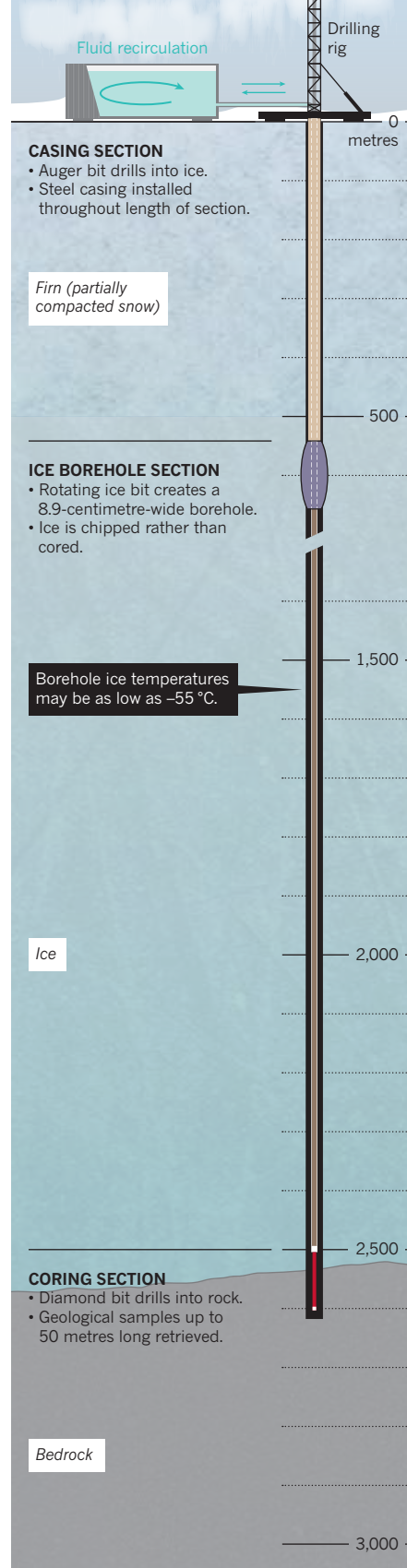
A fourth project, dubbed RADIX, would use a much narrower hole than the others — just 2 centimetres across — to bore up to 3 kilometres in a few days. RADIX has gone through limited testing in Greenland, says team leader Jakob Schwander, a climate scientist at the University of Bern.

No one knows exactly what these drills might encounter when they hit bottom. They might penetrate into pristine lakes below the ice sheet, which microbiologists could explore. Or they might reveal heat radiating upwards from the bedrock, melting the ice in ways that scientists had not expected.

“It's very multidisciplinary,” says John Goodge, a geologist at the University of Minnesota Duluth and a project leader on US RAID. “There's all kinds of stuff this fast technology allows that we were never able to have before.” ■

CLIMATE CLUES

The US-backed Rapid Access Ice Drill (RAID) seeks to penetrate into Antarctic ice faster than ever before.


MORE ONLINE

TOP STORIES

- Bronze Age skeletons were earliest plague victims go.nature.com/pd4hle
- Genetics probe identifies new Galapagos tortoise species go.nature.com/8vy6ga
- Dead star spotted eating planet leftovers go.nature.com/yflx5e

NATURE PODCAST

How cancers spread, bipolar disorder in the brain, and making carbon dioxide useful

nature.com/nature/podcast

ASTRONOMY

US grants trapped in vicious circle

Astronomers' resubmissions drive plunge in success rates.

BY CHRIS CESARE

Astronomers and astrophysicists in the United States are seeing their grant applications rejected at increasing rates because of stagnant budgets and an uptick in the number of resubmitted proposals, according to a draft report written for an advisory committee to the US National Science Foundation (NSF). The document, posted on the arXiv preprint server on 4 October, comes ahead of a November meeting set to discuss the issue (P. Cushman *et al.* Preprint at <http://arxiv.org/abs/1510.01647>; 2015).

The report highlights more than a decade of falling success rates for astronomical-science grants at the NSF and NASA as the number of proposals has increased faster than agency budgets. One key NSF programme in astronomy and astrophysics, for instance, funded fewer than 20% of proposals in 2014 — down from nearly 40% in 2002. And some NASA programmes saw rates fall from around 30% to 18% between 2004 and 2015.

The report rules out many explanations that scientists have suggested for the drop, such as a decrease in the quality of proposals; data from NASA show that, among the astrophysics grant proposals submitted to the agency, the fraction receiving scores from 'very good' to 'excellent' remained roughly constant from 2007 to 2012.

Instead, the report concludes, the main problem is that whereas funding has stayed flat, the total number of astronomers has continued to grow — and so the rate of resubmitted proposals has risen even faster because investigators who fail to secure funding in one year often try again the next. These resubmissions now account for a disproportionate number of grant applications, compounding the problem and leading to the dramatic drop in success rates.

The report enumerates the "knobs" that agencies can adjust to improve success rates, such as reducing the size of the average grant or shifting money from facilities to investigators — an idea that deserves a closer look, says Keivan Stassun, an astronomer at Vanderbilt University in Nashville, Tennessee, and a co-author of the analysis.

Other fixes, such as capping the number of proposals from investigators who have submitted too many unsuccessful applications, only "disguise the problem", the report argues. ■



Life Study aimed to find associations between factors early in life and outcomes later on.

EPIDEMIOLOGY

Massive UK baby study cancelled

After demise of similar US project, decision prompts rethink about design of future cradle-to-grave efforts.

BY HELEN PEARSON

An ambitious study that planned to collect information on 80,000 British babies throughout their lives has ended just 8 months after its official launch because not enough prospective parents signed up. The closure comes less than a year after the US National Institutes of Health (NIH) cancelled a similar effort to trace 100,000 children from birth, prompting fears that researchers will now shy away from proposing similar studies.

"I am afraid that the scientific community may not dare to embark on similarly ambitious cohort studies in the near future," says Camilla Stoltenberg, who heads the Norwegian Institute of Public Health in Oslo. She is responsible for a major birth-cohort study in Norway and chaired the international scientific-advisory committee to the UK project, called Life Study.

Prized by both medical researchers and social scientists, birth-cohort studies reveal associations between factors early in life, such as poverty or a mother's diet in pregnancy, and outcomes later on, ranging from diseases to cognition and earnings. Various efforts already exist around the world, but Life Study was to be one of the biggest and most ambitious yet. It got the green light in 2011 when government funding bodies, including the Economic and Social Research Council (ESRC) and the Medical Research Council, agreed to support the study with £38.4 million (US\$58.9 million) until 2019.

In January 2015, a team led by Carol Dezateaux, a paediatric epidemiologist at University College London's Institute of Child Health, opened the study's first dedicated recruitment centre, on the outskirts of London. The researchers hoped to sign up as many as 16,000 prospective mothers — of a total target of 60,000 — by July 2016. Another

MASTERFILE/CORBIS

20,000 babies were to be recruited nationwide after birth. But between January and early September this year, just 249 women signed up, according to the ESRC, which oversaw the study. A review of the project in July identified recruitment as a major concern, and on 10 July, the ESRC decided that the study should close. The cancellation was publicly announced on 22 October.

PREMATURE DEMISE?

Dezateux and some of her colleagues say that the closure was premature, and that they were not sufficiently consulted on the decision. They accept that recruitment was difficult, a challenge intensified by the study's remit to include a substantial proportion of families from ethnic-minority and disadvantaged groups, who have historically been particularly hard to recruit.

But the researchers say that they intended to test and refine recruitment methods during the first phase of the study — for example, the team had planned to make the study less burdensome for women by collecting information during a routine ultrasound scan rather than asking for a separate visit — and that the review process did not take such plans fully into account.

Fiona Armstrong, who was responsible for Life Study at the ESRC, says that the research council did indeed consider the researchers' plans to adjust the recruitment process — and consulted the research team as part of that

process — but ultimately, it still concluded that “whatever might be done wasn't enough”. “We couldn't take the risk of putting more and more money into it,” she says. The study consumed around £9 million (\$13.8 million), a sliver of the more than \$1.2 billion — over 15 years — that was sunk into the US National Children's Study (NCS).

Epidemiologists are drawing parallels between Life Study's demise and that of the NCS. “It's déjà vu all over again,” says Mark Klebanoff, a paediatric epidemiologist at Nationwide Children's Hospital in Columbus, Ohio.

Clinical epidemiologist George Davey Smith, who co-directs a separate birth-cohort study at the University of Bristol, UK, notes that a huge challenge for both efforts was that they were trying to provide answers to extremely diverse questions, which put

“We have to find ways of doing this that pose the least burden possible to participants.”

constraints on the studies' designs. For example, assessing inequalities between socio-economic groups requires data from a large, representative population sample that includes disadvantaged and minority groups — whereas answering questions relating to the origin of disease requires the collection of extensive biological samples such as blood and tissues. “It's incredibly sad,” he adds of Life Study's end.

Those involved in both studies hope to salvage something from the wreckage. After the NCS ended, plans emerged for a more modest study of influences on child health; Dezateux says that she and her colleagues are “determined to take forward key elements” of Life Study.

Whether and how such studies can be conducted in future is unclear. Response rates are falling in many surveys and population studies compared with those in decades past, say researchers — perhaps because there are more demands on people's attention. “We have to be mindful of the fact that people's lives are busier than ever,” says Klebanoff. “We have to find ways of doing this that pose the least burden possible to participants.”

Scientists need to exploit existing data sources more, says Stoltenberg. Extensive databases of health, educational and income data exist in many countries and provide vast amounts of information on the cheap — as long as people consent to their use. In Norway, such databases have been crucial to the success of its national birth-cohort study, which is following more than 100,000 children, she says.

But it is important to create systems through which information can be more easily extracted from such databases for use in cohort and other types of research, she adds. “We don't have the infrastructure,” she says. “We're trying to drive sophisticated vehicles like birth-cohort studies where there are no real roads.” ■

PLANETARY SCIENCE

Falling junk has scientific value

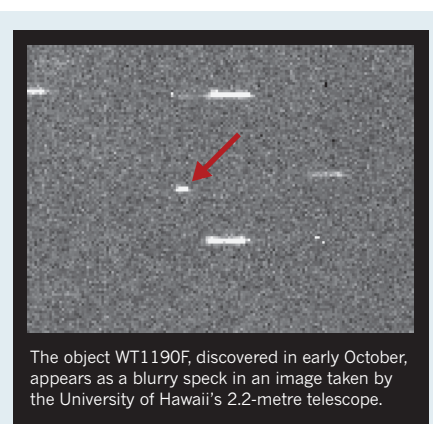
Astronomers prepare to observe an impact off Sri Lanka.

BY TRACI WATSON

Researchers call it sheer coincidence that a newly discovered piece of space junk is officially designated WT1190F. But the letters in the name, which form the acronym for an unprintable expression of bafflement, are an appropriate fit for an object that is as mysterious as it is unprecedented.

Scientists have worked out that WT1190F will plunge to Earth from above the Indian Ocean on 13 November, making it one of the very few space objects whose impact can be accurately predicted. More unusual still, WT1190F was a ‘lost’ piece of space debris orbiting far beyond the Moon, ignored and unidentified, before being glimpsed by a telescope in early October.

An observing campaign is now taking shape to follow the object as it dives through Earth's atmosphere, says Gerhard Drolshagen,



co-manager in Noordwijk, the Netherlands, of the European Space Agency's near-Earth objects office. The event not only offers a scientific opportunity to watch an object plunge through the atmosphere, but also tests

the plans that astronomers have put in place to coordinate their efforts when a potentially dangerous space object shows up. “What we planned to do seems to work,” Drolshagen says. “But it's still three weeks to go.” ▶

► WT1190F was detected by the Catalina Sky Survey, a programme based at the University of Arizona, Tucson, aimed at discovering asteroids and comets that swing close to Earth. At first, scientists didn't know what to make of this weird body. But they quickly computed its trajectory after collecting further observations and unearthing 2012 and 2013 sightings from telescope archives, says independent astronomy-software developer Bill Gray, who has been tracking the debris with astronomers at NASA's Jet Propulsion Laboratory in Pasadena, California.

WT1190F travels in a highly elliptical orbit, swinging out twice as far as the Earth–Moon distance, Gray says. His calculations show that it will hit Earth at 06:20 UTC, entering the ocean about 65 kilometres off the southern tip of Sri Lanka (see 'Splashdown'). Much, if not all, of it will burn up in the atmosphere, but "I would not necessarily want to be going fishing directly underneath it", Gray says.

The object is only 1 to 2 metres in size, and its trajectory shows that it has a low density, and is perhaps hollow. That suggests an artificial object — "a lost piece of space history that's come back to haunt us," says Jonathan McDowell, an astrophysicist at the Harvard–Smithsonian Center for Astrophysics in Cambridge, Massachusetts. It could be a spent rocket stage or panelling shed by a recent Moon mission. It is also possible that the debris dates back decades, perhaps even to the Apollo era. An object seen orbiting Earth in 2002 was eventually identified as a discarded segment of the *Saturn V* rocket that launched the second mission to put humans on the Moon.

WT1190F is a rare breed of space object. Researchers are currently tracking only 20 or so artificial objects in distant orbits, says Gareth Williams, an astronomer at the Minor Planet Center in Cambridge, Massachusetts. There are probably many more such pieces of space junk in orbit around the Earth–Moon system, but it is impossible to say how many. No others are known to have made the return trip to Earth, although it is likely that some have done so without anyone noticing, McDowell says.

Drolshagen plans to get spectral information on the object, which may help to identify it, and he hopes to coordinate impact observations conducted on ships or aeroplanes. But that may be the end of the concerted effort to study this class of object. Unlike near-Earth asteroids, space debris that flies well away from Earth has not commanded significant amounts of funding or attention. The US military, which tracks space debris, says that it lacks the ability to identify WT1190F or to predict its path.

"There is no official, funded effort to do tracking of deep-Earth orbits the way we track low-Earth orbit," McDowell says. "I think that has to change". ■

ONCOLOGY

Cancer-fighting viruses near market

Anticipated approval in Europe and the United States could spur a promising field with a chequered past.

BY HEIDI LEDFORD

An engineered herpesvirus that provokes an immune response against cancer seems poised to become the first treatment of its kind approved for use in Europe and the United States. On 23 October, advisers to the European Medicines Agency endorsed the approval of a genetically engineered virus called talimogene laherparepvec (T-VEC) to treat advanced melanoma. In April, advisers to the US Food and Drug Administration (FDA) did the same, and the agency is expected to approve T-VEC this month.

With dozens of ongoing clinical trials of similar 'oncolytic' viruses, researchers hope that such an approval could generate the enthusiasm and cash needed to spur further development of the approach. "The era of the oncolytic virus is probably here," says Stephen Russell, a cancer researcher and haematologist at the Mayo Clinic in Rochester, Minnesota. "I expect to see a great deal happening over the next few years."

Many viruses preferentially infect cancer cells. Malignancy can suppress normal antiviral responses, and sometimes the mutations that drive tumour growth also make cells more susceptible to infection. Viral infection can thus ravage a tumour while leaving abutting healthy cells untouched, says Brad Thompson, president of the pharmaceutical-development firm Oncolytics Biotech in Calgary, Canada.

EARLY ATTEMPTS

The strategy builds on a phenomenon that has been appreciated for more than a century. Physicians in the 1800s noted that their cancer patients sometimes unexpectedly went into remission after experiencing a viral infection. These case reports later inspired doctors, particularly in the 1950s and 1960s, to raid nature's viral cupboard. Clinicians injected cancer patients with a menagerie of viruses. Sometimes the therapy

destroyed the tumour, and sometimes it killed the person instead.

Unlike the wild viruses used in those mid-twentieth-century experiments, some of today's anti-cancer viruses are painstakingly engineered. T-VEC, for example, has been altered to drastically reduce its ability to cause herpes. Researchers also inserted a gene encoding a protein that stimulates the immune system, which makes the virus even more potent against cancer (see 'Going viral against cancer').

As more researchers entered the field and initiated small clinical tests, they began to produce enticing anecdotes. Russell recalls the case of an individual with myeloma who remained sick after undergoing two stem-cell transplants. A tumour on the left side of her forehead had degraded the bone underneath and was putting pressure on her brain. Yet treatment with an experimental virus sent her into complete remission (S. Russell *et al. Mayo Clin. Proc.* **89**, 926–933; 2014). "She's a star patient who convinced us that this oncolytic paradigm can really work," he says.

But statistics — not anecdotes — rule over drug approvals. In 2005, regulators in China approved an oncolytic adenovirus called H101 to treat head-and-neck cancer, after evidence showed that the treatment could shrink tumours. Those trials stopped short of assessing improvements in patient survival — a measure often required for FDA approval. Since then, a medical-tourism industry has built up in China for people who cannot get the therapy in their home countries.

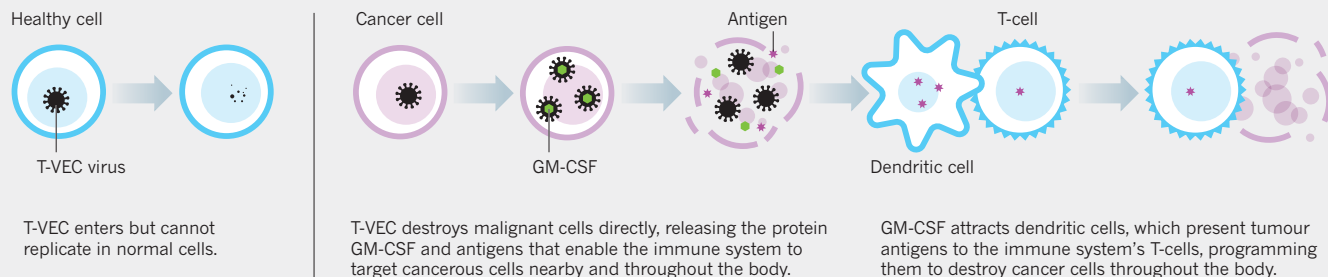
Then, in May this year, a team supported by biotechnology giant Amgen of Thousand Oaks, California, published promising results from a large clinical trial of T-VEC (R. H. Andtbacka *et al. J. Clin. Oncol.* **33**, 2780–2788; 2015). The virus both shrank tumours in people with advanced melanoma and extended patient survival by a median of 4.4 months. Yet statistically, survival benefits fell just a hair's breadth of significance. "That raised the question, 'Well, what is statistical significance? Is this an active agent or not?'" Russell says.

He and others note that the therapy — which must be injected directly into tumours — seemed to rein in cancer elsewhere in the

Viral infection can ravage a tumour while leaving abutting healthy cells untouched.

GOING VIRAL AGAINST CANCER

The virus-based cancer therapy T-VEC infects tumour cells and destroys them by stimulating the immune system to direct an attack against malignant cells in the body.



body as well. This is a sign that results are real and that the virus sparked an immune response as intended, Thompson says.

ROOM FOR IMPROVEMENT

Administering T-VEC in combination with cancer immunotherapy could prove particularly effective, notes Stephen Hodi, an oncologist at the Dana-Farber Cancer Institute in Boston, Massachusetts. In June 2014, a small clinical trial by Amgen suggested that this combination may boost effectiveness over that of the immunotherapies alone.

And researchers continue to look for ways

to improve T-VEC. In particular, they would like to be able to deliver the therapy systemically, so that the virus could target tumours in organs that are difficult to reach with an injection. This would require a technique to prevent the body from mounting an immune response to the virus prematurely, which would disable it before it could reach and kill tumour cells, says Howard Kaufman, a cancer researcher at Rutgers Cancer Institute of New Jersey.

To that end, those in the field are experimenting with a smorgasbord of viruses — from poxviruses to vesicular stomatitis virus, which does not normally infect humans but causes a

blistering disease in cattle. Oncolytics Biotech is studying a virus that hitch-hikes through the body on certain blood cells, camouflaged from the immune system.

If cancer-killing viruses could be delivered to their targets through the bloodstream, rather than via injection directly into the tumour, they could be used to treat a greater range of cancers. Thompson envisions a day when physicians will be able to peruse a menu of oncolytic viruses and select the best fit. "Each virus interacts with the immune system differently," he says. "They could have a role in pretty much all cancer therapy." ■



Residents of Kiribati make their way home through floodwaters. Rain and severe weather can exacerbate the high tides that often rush into villages.

Before we drown we may die of thirst

The island nation of Kiribati is one of the world's most vulnerable to rising sea levels. But residents may have to leave well before the ocean claims their homes.

BY KENNETH R. WEISS

High tide left its mark on the houses like a dirty ring in a bathtub. The flood crept into the village of Teaoraereke under the cover of darkness, sending filthy seawater sloshing through pigsties and shallow graves, and into people's homes.

CIRIL JAZBEC

Teaoraereke residents scrambled to retreat, hoisting sleeping children, sodden bedding and other belongings to higher ground. But some stayed put, including Rerema Kauria, a 63-year-old grandmother who was marooned just inches above the floodwaters on a raised platform bed. She was still there by mid-morning as the water receded, her possessions tucked into the rafters of her traditional house of wooden poles and thatch. She knew that when high tide returned that afternoon it would bring more flooding, but she gave a roaring laugh when asked if she had considered leaving. "Where would I go?"

The uncertain future of people such as Kauria has drawn attention to a collection of atolls in the central Pacific Ocean that make up the Republic of Kiribati (pronounced Keer-re-bahs). The average height of the country's 33 islands is little more than 2 metres above the ocean,

which makes Kiribati acutely vulnerable to climate change. By the end of the century, melting polar ice and the thermal expansion of warmer seawater is expected to raise global ocean levels by perhaps 1 metre. That upsurge would, according to some predictions, displace many from Kiribati and millions of others around the world — and the water will keep going up.

For years, Kiribati President Anote Tong has sounded the alarm over his nation's plight, warning that residents would soon have to abandon their homeland. The flooding that hit Teoraereke last year reinforces those dire predictions. Although it is impossible to know how much, if at all, climate change contributed to the flooding, village residents say that they have never before seen such inundation. To some of them, it seemed as if the swelling seas were starting to consume Kiribati and the end of the atoll might come sooner than they had thought.

But researchers who study Kiribati say that the situation is not a simple story of rising seas swallowing low-lying islands. In fact, some coastal experts dispute the idea that Kiribati will soon sink beneath the waves like a modern Atlantis. They have gathered evidence that many of these islands have been gaining ground in recent decades by capturing sediments from surrounding coral reefs. "It's just plain wrong to assume that all atolls are washing away," says Arthur Webb, a coastal geomorphologist affiliated with the University of Wollongong in Australia who has spent two decades living and working in the Pacific Islands. "It's also wrong to sugar-coat the sobering facts that rising sea levels will ultimately seal the fate of low-lying islands and their limited soils and groundwater. The confusion isn't surprising. It's just more complicated than many expect."

Even if Kiribati does not drown in the near future, its residents may soon need an exit strategy. Poverty, overcrowding and poor sanitation are galloping ahead of rising seas to deplete the islands' resources, especially their supply of clean fresh water. And residents' habits of altering the shoreline and removing coastal protections can magnify the impacts of the swelling oceans, leaving villages more exposed to flooding.

The story playing out on these tiny islands shows how difficult it is to tease out the impact of climate change from other human and environmental pressures. And what happens to the Kiribati people has implications for the hundreds of millions in low-lying coastal areas across the globe who will be threatened with flooding and displacement in coming decades. But unlike the residents of Miami, Guangzhou or Mumbai, the Kiribati people have no option of retreating inland or up-slope as their vulnerable flyspecks of land become uninhabitable. As Kauria says: where would they go?

AT THE MERCY OF THE TIDES

From the air, Kiribati's Tarawa atoll emerges from the Pacific as narrow strands of land that join to form a wispy V shape. On the outside of the V is the deep blue of the ocean; inside are the aquamarine and turquoise waters of the shallow lagoon.

Tarawa is the capital of Kiribati, which is one of the most remote countries on Earth, located on the equator about halfway between Australia and Hawaii. Its atolls are scattered across a patch of the Pacific the size of India, and yet they have a total of just 811 square kilometres of land, about half the size of Greater London.

When a plane lands on Tarawa, a crowd gathers at the airport, drawn by the excitement of the jet making the three-hour flight from Fiji. Aside from occasional freighters bringing canned food, this twice-weekly Fiji Airways flight provides the primary connection to the outside world.

The airport was built on relatively high ground, an elevation of 3 metres, in one of the atoll's widest sections. It happens to sit above the main subsurface reservoir, a freshwater layer floating on top of the seawater that presses against the porous island from all sides.

Although the sea presents an existential threat, the more immediate problem is not too much water, but too little — of the fresh, clean kind. A dozen of Kiribati's islands are deserted, too arid to support human

habitation. Without enough replenishing rains, their thin lenses of groundwater turned brackish. On Tarawa, groundwater is heavily overdrawn and contaminated by the local practice of defecating on the beach or in the bushes. With little land, residents bury their relatives and raise pigs next to their homes, which also contributes to groundwater pollution.

Next to the road leading to the airport, a buried white plastic pipe that carries fresh water from the reservoir has been exposed in places, owing to erosion by waves and tides. Public workers have fought back by using old tyres filled with concrete to hold it in place. They have had less success keeping locals from tapping illegally into the waterline, directing the flow into hand-dug wells for their homes. Water supplies are so limited that authorities turn on the airport's groundwater pumps for only a couple of hours every other day.

The perils of water, both sweet and salty, are intertwined with Tarawa's history, says George Fraser, high commissioner to Kiribati from Australia, which is the biggest provider of international support to the developing nation. In one of the bloodier clashes in the Pacific during the Second World War, US commanders misjudged the tides and landing craft got stuck on the reef, forcing marines to wade through chest-deep water under heavy Japanese fire.

Fraser deconstructs that infamous battle as he takes a tour of the island the day before the king tide that flooded Teoraereke. A fast and confident driver, Fraser weaves his small sport utility vehicle around wobbly wheeled trucks, dodging potholes. In narrow spots, the atoll's main road is soaked with seawater and he swerves to avoid a wave splashing over a concrete berm. "Some people use calendars to get through the week," he says. "We use tide charts."

The road is the only paved one on the atoll. It crosses a series of causeways that have been battered by wheels and waves; road crews repair cavities, stuffing them with as much concrete and sand as possible to slow the decay. The Australian government has bankrolled much of a repaving project along the length of South Tarawa.

As Australia's top representative, Fraser has a keen sense of the various challenges that this poor country faces in coming decades, and how they stack up. "If you look at rising sea levels as the train coming down the track, it's a couple of kilometres away," he says. "If you look at what's 100 metres down the track, it's no water, and right behind it is no food."

More than half of Kiribati's 110,000 residents live on Tarawa, and their numbers are rapidly increasing as more arrive from outer islands seeking jobs, cash and better schools. Many were subsistence fishers and farmers on their home islands, struggling with depleted fisheries and poor soil damaged by periodic over-wash of salt water. When they get to Tarawa, they often end up

jobless or underemployed.

The Kiribati culture is communal, with families accustomed to bedding down together on woven mats on the floor. It is taboo to refuse the request of a relative, so households often pack dozens of extended family members from other islands under one roof. That has made South Tarawa one of the most densely packed places in the Pacific; its clusters of shanties resemble slums in the poorest capitals of Africa and Asia. Factoring in high birth rates and ongoing urbanization, the government projects that the population of the island will almost double in 15 years. The new Battle of Tarawa will be over where all these people will live.

THE INCREDIBLE SHRINKING ISLAND

Today's scientific debate about whether Kiribati is growing or shrinking can be traced to Charles Darwin — who first worked out how coral atolls form. While sailing the Pacific on the HMS *Beagle* in the 1830s, he theorized that these curiously shaped sand islands are produced by coral reefs that sprouted on the slopes of volcanic islands and have continued to grow as the volcanoes sink into the abyss. He was proved right a century later, when scientists drilled into an atoll and hit volcanic rock.

"SOME PEOPLE USE CALENDARS
TO GET THROUGH THE WEEK.
WE USE TIDE CHARTS."

Over the millennia, the exoskeletons of millions of tiny coral animals fuse with coralline algae and the shells of molluscs and other sea creatures to form limestone reefs, often arranged in a circle with a shallow lagoon in the middle. Living corals grow on the fringes of these limestone platforms. As the crest of the living reef reaches close to the ocean surface, waves break some of it into rubble and sand that gets deposited on the dead limestone platform to form land.

The atolls that exist today are the survivors, ones in which coral reefs kept pace with rising seas and the subsidence of the undersea volcano. The pressing issue is, what will become of those atolls as sea levels start rising faster? Researchers wonder whether corals can keep up, given the host of environmental problems they face. In many places, overfishing and nutrient pollution have triggered the growth of coral-killing bacteria and algae. Abnormally warm seawater is causing 'bleaching' die-offs throughout the tropics, and as ocean water takes up more carbon dioxide and acidifies, it will be harder for coral polyps to build rugged exoskeletons.

Around Tarawa, the coral reefs are in particularly poor shape, says Simon Donner, a climatologist at the University of British Columbia in Vancouver, Canada, who has done diving surveys. "Coral cover is lower than you'd expect around the island," he says. "That's the legacy of pollution, sewage mostly, and frequent bleaching events in the past 20 years."

To help predict how corals may fare in the future, Dennis Hubbard, a geologist at Oberlin college in Ohio, and his colleagues have been peering into the past, amassing a database of sediment core samples obtained by drilling into limestone beneath coral reefs. With carbon dating, they can determine how quickly these reefs have grown: in yet-to-be-published work, they have found that more than half of the world's coral reefs grew more slowly over the past 10,000 years than sea levels are rising today. Extrapolating forward, those results suggest that only half of all atolls in existence today have a chance of keeping pace with rising seas under the best of conditions, he says. "Given that this was in a time with no human impact, we feel this is the most optimistic scenario possible."

NOT SO FAST

Kiribati and other low-lying island nations have long been held up as the countries most susceptible to the ravages of rising seas. In 2001, the Intergovernmental Panel on Climate Change (IPCC) highlighted predictions that two-thirds of Kiribati and the nearby Marshall Islands would be inundated by a sea rise of 80 centimetres.

But the idea that these atolls will disappear any time soon has been challenged by Paul Kench, a coastal geomorphologist at the University of Auckland. He and his colleagues have pored over satellite images, comparing new and old aerial photographs to see how such islands have changed.

In a 2010 study, he and Webb determined² that 23 out of 27 atoll islands scattered across Kiribati, Tuvalu and the Federated States of Micronesia had either increased in area or remained stable in recent decades. The results, they reported, "contradict widespread perceptions that all reef islands are eroding in response to recent sea level rise". The researchers concluded that these islands are more "resilient landforms" than previously thought. The study created a media stir in the region and beyond. It has been widely cited by climate-change sceptics seeking to punch holes in research on global warming and its impacts.

Kench recognizes the powerful forces of climate change but complains that too many scientists and activists focus solely on rising sea levels while ignoring the other part of the equation: how the land responds. "There are a lot of claims that islands are passive geological entities that will sit there and drown," he says. "Our work shows that they are anything but static. They are dynamic. They move around and they can grow. So just because sea level is rising, it doesn't mean doom and gloom for all atolls."

He also believes that most scientists make a mistake by tethering the fate of atoll islands to the health of surrounding coral reefs. Even if reefs

"IF WE RETREAT FROM THE OCEAN SIDE, AND INSTITUTE A SETBACK, WE WILL FALL INTO THE LAGOON."



A man rebuilds the sea wall that protects his home on South Tarawa.

die, he says, they can provide sufficient sediment to maintain islands for a century or more.

But Hubbard considers Kench's views shortsighted. "If you run out of reefs, you run out of sediment, and once you run out of sediment, you run out of islands," he says. "A lot of this is a semantics issue, challenging when the reef island is going to be physically underwater. Those reef islands are going to be abandoned long before that because they are uninhabitable."

On Tarawa and other Kiribati islands (see 'Isolated islands'), most people do not dwell on such matters, going about their daily lives just like residents of other countries. But their president has earned international recognition for speaking out on the threats of climate change.

In an interview, Tong dismisses those who suggest that atolls are resilient to rising seas, saying that they have the luxury of "talking from the top of a mountain" and not putting their lives on the line. "These people are not living here. Their grandchildren will not be living here. If they believe that, let them come here," he says, pounding his fist on a chair armrest for emphasis. "I'd rather plan for the worst and hope for the best."

Tong has told his people that they must prepare to leave, seeding the idea of an early "migration with dignity", rather than fleeing as refugees when storm-generated waves wash

over the islands. Last year, his government completed an US\$8-million purchase of 22 square kilometres of hilly land in Fiji, to grow food and provide possible refuge for some of his people — although it will not accommodate all of them. He does not know when people will need to migrate, but he wants to purchase more land in Australia and New Zealand, saying that it is much cheaper than trying to build sea walls and other defences. "If we build up these lands, it's going to cost billions of dollars," he says. "We might as well be buying land for millions of dollars elsewhere."

DEVELOPING CHALLENGES

For nearly a week after Teoraereke flooded, resident Matua Kamori worked alongside his neighbours to build a makeshift sea wall where the high tide had breached a sand berm on the beach. Villagers piled up chunks of coral scavenged from the shore and grouted them together

JUSTIN MCMAHUS/THE AGE/FAIRFAX MEDIA/GETTY



Children play in a king tide that flooded their family's land on Tarawa.



with cement donated by a local church.

Kamori, 33, lives in the village with his wife and 4 kids on a small parcel of land given to him by his wife's uncle in exchange for looking after one of his sons. To prepare the land, Kamori spent months scouring sand and coral gravel off the beach and hauling it to the site with nothing more than a rice sack. Over time, he fashioned a building pad half a metre high and constructed a hand-hewn traditional house of wood and thatch on it. Such beach mining is rampant on Tarawa, according to household surveys. Government studies³ show that it increases the likelihood of flooding by lowering the protective sand berm that keeps the highest tides at bay.

In the case of the recent inundation, Kamori says that he fared better than most: the water reached calf-deep in his house, rather than thigh-high. But nothing could be done to stop the briny stew of salt water, mixed with human and animal waste, from polluting his well or killing his garden of vegetables and banana and papaya trees.

Kamori says that he settled on the land because he had nowhere else to go. As crowding increases, new settlements are pushing into vulnerable lowlands, places they historically would have avoided.

The individual actions of settlers such as Kamori are only part of the problem on Tarawa. Large-scale construction projects over the years have also exacerbated flooding and erosion, says Naomi Biribo, Kiribati's secretary of fisheries and marine resources development. Biribo earned a PhD in Australia by examining the impact of the sea walls and other human structures on Tarawa. The construction of causeways, rather than bridges, to connect the islets had the effect of closing channels and disrupting the flow of sediment that normally resupply some eroding coastlines, she found⁴. Reclamation projects that create new land are another problem: although such efforts have added hundreds of hectares to Tarawa, they accelerate erosion elsewhere, says Biribo.

For Tarawa residents, she says, the thin ribbons of land leave little room to move. "In many places on Tarawa, you can stand in the middle and you can see the ocean on one side and still see lagoon on the other side," she says. "If we retreat from the ocean side, and institute a setback, we will fall into the lagoon."

Biribo's work suggests that sea-level rise may be having a small influence on the shoreline changes happening today, but nothing compared to human activities and the seasonal variations in erosive tides and waves in the Pacific that come with the El Niño periods of warming and La Niña cooling.

Donner agrees that climate change has been dwarfed by other factors so far. "You cannot blame the flooding on sea-level rise," he says. "At least not yet."

Where does this leave residents of Kiribati? Webb has long wrestled with that question. He is married to a woman from Tarawa, and they own a house there, where they live with their children for part of the year.

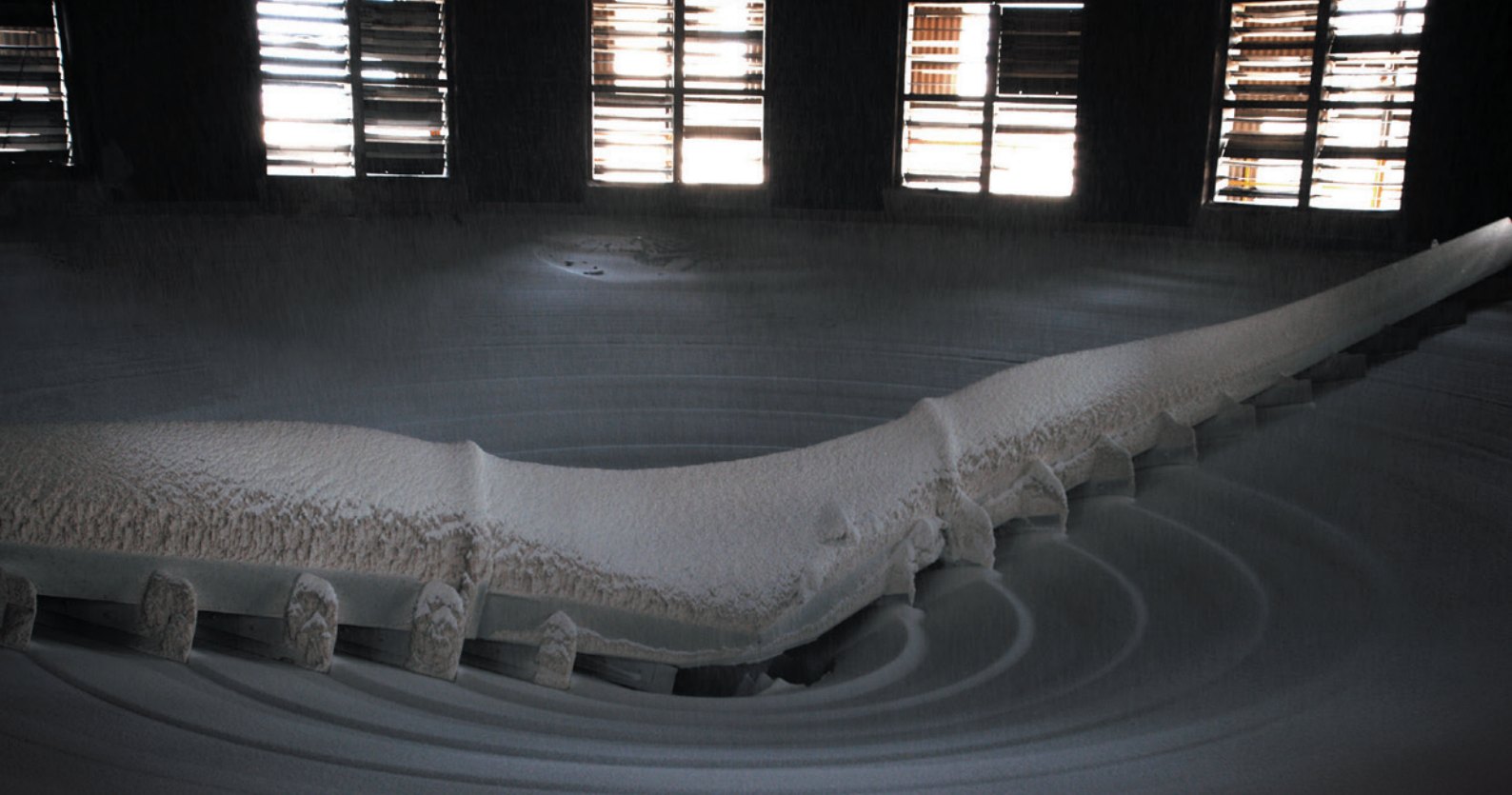
Webb was also a lead author of the small-islands chapter⁵ of the IPCC's fifth assessment report in 2014, which found that rising seas present "severe sea flood and erosion risks for low-lying coastal areas and atoll islands". It highlighted one projection⁶ that a 50-centimetre rise in sea level could displace 1.2 million people from low-lying islands in the Caribbean Sea and the Indian and Pacific oceans; that number almost doubles if the sea level rises by 2 metres. And yet, the latest assessment steered clear of the IPCC's previous assertion that an 80-centimetre rise would inundate two-thirds of Kiribati.

Scientific understanding of atoll geology has sharpened since that earlier projection. Webb expects some remnants of Tarawa to remain a century or two from now, but probably no more than some wave-washed gravel banks — and by that point, everyone will have long gone.

The geological evidence does not get to the key human question about the destiny of these Pacific islanders. That leaves Webb facing a difficult question — one he hears from his own Kiribati-born teenagers. "How long do we have?" they ask. To that, he replies: "Your children will not grow old in the atolls." ■

Kenneth R. Weiss is a freelance writer in Carpinteria, California. A grant from the Pulitzer Center on Crisis Reporting contributed to research for this article.

1. Smith, J. B. et al. In: *Climate Change 2001: Impacts, Adaptation, and Vulnerability. Contribution of Working Group II to the Third Assessment Report of the Intergovernmental Panel on Climate Change* (eds McCarthy, J. J., Canziani, O. F., Leary, N. A., Dokken, D. J. & White, K. S.) Ch. 19 (Cambridge Univ. Press, 2001).
2. Webb, A. P. & Kench, P. S. *Global Planet. Change* **72**, 234–246 (2010).
3. Webb, A. P. *Technical Report – An Assessment of Coastal Processes, Impacts, Erosion Mitigation Options and Beach Mining* EU EDF–SOPAC Project Report 46 (SOPAC, 2005).
4. Biribo, N. & Woodruff, C. D. *Sustain. Sci.* **8**, 345–362 (2013).
5. Nurse, L. A. et al. In: *Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part B: Regional Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Barros, V. R. et al.) Ch. 29 (Cambridge Univ. Press, 2014).
6. Nicholls, R. J. et al. *Phil. Trans. R. Soc. A* **369**, 161–181 (2011).



HOW TO MAKE THE MOST OF CARBON DIOXIDE

Researchers hope to show that using the gas as a raw material could make an impact on climate change.

BY XIAOZHI LIM

On 29 September, the XPRIZE Foundation based in Culver City, California, announced a 4½-year competition that will award US\$20 million to the research team that can come up with the best way to turn carbon dioxide from a liability into an asset.

With gigatonnes of the gas pouring into the atmosphere each year, and with the consequences for global climate becoming increasingly obvious, the Carbon XPRIZE would reward technologies that can convert CO₂ emissions from coal and natural-gas power plants into useful products such as alternative building materials, fuels and raw material for the manufacture of plastics and other chemicals.

The invitation should have plenty of takers: a growing number of companies and research chemists are already pursuing that goal. In Reykjavik, what looks like an overgrown playground climbing frame is actually a small chemical plant that turns CO₂ into methanol: a fuel that can also be used to manufacture products ranging from paints to wrinkle-resistant textiles. In Houston, Texas, another small plant is turning CO₂ into materials that go

on to become coatings and adhesives. And in Tokyo, Japan, Asahi Kasei Chemicals is widely licensing its technique for turning CO₂ into the polycarbonate plastics used in bulletproof glass, spectacle lenses and electronic parts.

Using this greenhouse gas as a raw material is an idea that many scientists once dismissed as hopeless, says Chunshan Song, a chemical engineer at Pennsylvania State University, University Park. As a practical matter, he says, “lots of people believed that nothing could be done with CO₂ utilization” after the stuff went up the smokestack.

As a source of carbon, sceptics argued, the gas was far more difficult and expensive to obtain than the petroleum, coal and natural gas that now provide the raw material for most chemical manufacturing. And even if CO₂ could be captured cheaply enough, converting such a stable molecule into more-useful chemicals would generally require lots of energy, which might well come from fossil-fuel plants. The conversion could cost a fortune and make more CO₂ than it consumed.

But the balance is starting to shift. New conversion technologies are allowing

ASIM HATEEZ/BLOOMBERG/GETTY

Pellets of urea fertilizer are made from carbon dioxide in a plant in Daharki, Pakistan.

the energy-hungry chemical reactions to proceed more efficiently. Renewable sources such as solar and wind can increasingly supply that energy at a competitive cost without the carbon penalty. And solutions to the capture problem may be forthcoming if governments follow the recommendations of the Intergovernmental Panel on Climate Change (IPCC) and mandate carbon capture and storage — grabbing the CO₂ coming out of power plants and other industries, and locking it away underground. But rather than simply burying it, companies could defray the cost of capture by putting the gas to productive use (see *Nature* **526**, 306–307; 2015).

Michele Aresta, a chemist at the University of Bari in Italy, estimates that if currently known processes were deployed most efficiently and at the greatest possible scale, they could directly use some 300 million tonnes of CO₂ per year, while indirectly reducing emissions by around a gigatonne per year — roughly 5% of the total net emissions¹. This is hardly a total solution to the CO₂ problem, he says, but it is substantial nonetheless. And that is not counting the additional benefits, which include a shift away from relying on fossil fuels as a source of carbon for materials and chemicals, as well as the use of CO₂-derived fuels to store and transport energy from wind, solar and other intermittent sources.

The Carbon XPRIZE may well accelerate these developments. But research into CO₂ conversion has already been receiving support from funding agencies worldwide. Since 2009, for example, the German Ministry of Education and Research has pumped in around €100 million (US\$110 million) to support research on the use of CO₂ in chemical and synthetic-fuel production. In 2011, the US Department of Energy invested \$106 million in various CO₂-utilization projects. The European Union has lined up a 2020 Horizon Prize worth €1.5 million for a technology that demonstrates viable CO₂ reuse. And over the next five years, China is expected to invest some 30 billion yuan (\$4.7 billion) in CO₂ recycling in the coal, steel, cement and paper industries.

LOW-HANGING FRUIT

Industrial uses of CO₂ have been around for generations. Every year, about 20 million tonnes of the gas are used 'as is': for example, it's the 'fizz' in fizzy drinks, and it can also be used as a high-pressure solvent to clean electronics. But gas used in this way quickly returns to the atmosphere, where it contributes to global warming.

Another 114 million tonnes of CO₂ go into the production of urea fertilizer every year, accounting for more than 60% of the total worldwide consumption of the gas. But urea is usually made by reacting CO₂ with ammonia — most of which is made with hydrogen

derived from fossil fuels. Researchers such as Peter Styring, a chemical engineer at the University of Sheffield, UK, are working with industrial partners to develop catalysts and reaction conditions that will allow the production of urea directly from nitrogen, water and CO₂. But currently, the urea industry is still a net source of CO₂.

In the quest to make more materials from CO₂ without adding to the greenhouse problem, chemists' first challenge is to overcome the molecule's stability. Unlike a carbon atom in isolation, which has four electrons that are available to form bonds, the carbon in CO₂ has given up all four to the tight grip of the oxygen atoms. That grip has to be loosened before

CHEMISTS' FIRST CHALLENGE IS TO OVERCOME THE MOLECULE'S STABILITY.

the carbon can form new bonds — and that takes energy (see 'The Carbon Challenge'). "If it [energy] comes from fossil fuels, it's pointless," says Styring. "Everything has to be from renewable energy."

This energy cost of separating the carbon and oxygen atoms is so exorbitant that some researchers in the field simply sidestep the problem. They focus on products such as calcium carbonate (CaCO₃), which can be made by combining CO₂ with calcium compounds in low-energy reactions that leave the carbon-oxygen bonds relatively undisturbed. Calcium carbonate also happens to have a large potential market: it is used as a construction material and for whitening factory-produced paper. And because it is more inert than CO₂, it could be used to store the gas over long time scales — something that already happens naturally in limestone deposits and coral reefs. "Nature has used mineralization to store millions of billions of tonnes of CO₂," says Martin Devenney, chief operating officer at Calera, a green-energy company in Los Gatos, California.

Most of the calcium carbonate now used in industry, an estimated 15 billion tonnes per year, comes from mining such deposits. But since 2013, Calera has been operating a demonstration facility that produces the material from CO₂ and carbide residue: an industrial waste that comes from the production of polyvinyl chloride (PVC). Calera first adds water to the residue to extract calcium hydroxide, then bubbles CO₂-rich flue gas from a nearby industrial plant through the solution to obtain pure calcium carbonate, which is turned into fibre cement boards that can be used in construction.

After factoring in energy costs, says Devenney, the company's process captures about 170 kilograms of CO₂ for every tonne of

calcium carbonate manufactured — versus zero kilograms for every tonne that is mined.

In Asia, meanwhile, manufacturers licensing technology developed by Asahi Kasei Chemicals are using CO₂ to produce about 660,000 tonnes of polycarbonates annually, roughly 14% of the global total. Polycarbonates are plastics used in products ranging from water bottles to spectacle lenses. They are usually made in a reaction involving phosgene (COCl₂): a toxic compound that is infamous for its use as a poison gas in the First World War. It is typically produced by combining chlorine with carbon monoxide made by burning coal. The Asahi Kasei process replaces phosgene with CO₂ — overcoming the

molecule's stability by putting it through three intermediate transformations.

Asahi Kasei's process is still unable to make polycarbonate chains that are long enough for applications such as water bottles. But it is cost-competitive with the standard method, and according to Aresta's estimates, reduces indirect CO₂ emissions from 6 tonnes for every tonne of polycarbonate produced to 1 tonne.

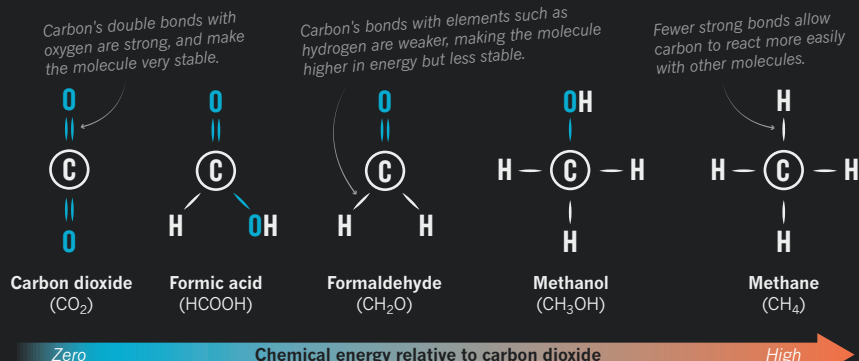
Yet another promising area for low-energy CO₂ utilization is in the production of polyols: a group of sugar-like molecules that are often used as the raw material for polymers such as polyurethanes, which are then used in mattresses, adhesives, coatings, refrigerator insulation and Spandex. Polyols are mostly made from precursor molecules known as epoxides. But Novomer, a green-chemistry company in Waltham, Massachusetts, has developed a cobalt-based catalyst that allows half of the epoxides to be replaced with CO₂. "Our catalyst will not work without the carbon dioxide," says Peter Shepard, executive vice-president of Novomer. The Novomer process is currently being tested at a plant in Houston, he says, where it is producing 3,000–4,000 tonnes of polyols per year while generating about one-third of the CO₂ that would be produced through conventional methods. And the company has begun designing a facility with a capacity of 100,000 tonnes per year, scheduled for completion in 2017.

AN UPHILL TASK

Despite these successes, the demand for low-energy products such as polycarbonates and inorganic carbonates is relatively limited. The big money is in hydrocarbons, which consist of one or more carbon atoms, often arranged in long chains or complex rings, surrounded by

THE CARBON CHALLENGE

Turning carbon dioxide into more-useful chemicals requires an input of energy to break its strong bonds. Each one-carbon molecule in this sequence has a higher chemical energy than the one before.



hydrogen atoms. As well as their many other uses, hydrocarbons are ubiquitous in petrol, diesel, jet fuel and just about every other liquid energy source on Earth — a market roughly 14 times larger than that for non-fuel chemicals.

Making hydrocarbons out of CO₂ is a “completely different story” from making low-energy products, says Aresta. In addition to the energy required to break the carbon atom free from oxygen so that it can link up with other carbons, the process needs a source of cheap hydrogen — which currently comes from fossil

As with other efforts to make fuels from CO₂, however, the challenge is to do it efficiently: even in the laboratory, the known methods consume far more energy than the resulting fuels can provide. Among the most vexing problems is how to obtain the hydrogen. Right now, the choices are to extract the gas from fossil fuels, a source that is cheap but hardly green, or to make it by splitting water, a process that is both energy intensive and costly.

“We need at least 15 years of good research”

MAKING HYDROCARBONS OUT OF CO₂ IS A COMPLETELY DIFFERENT STORY.

fuels. And it is not easy to control the reactions to get a target hydrocarbon molecule instead of a related molecule that is one carbon shorter or longer. “This is tricky,” says Song.

Still, researchers are making progress. Since 2010, for example, Song has developed a series of catalysts based on iron and cobalt that can bring CO₂ molecules close together with hydrogen molecules at 300 °C and about 10 atmospheres, forming ethylene, propylene and butane². Because his experiments are still being done in the laboratory, Song cannot yet say how much his process could reduce greenhouse-gas emissions when carried out on an industrial scale. But any headway could have a large effect: these three chemicals are among the most heavily used in the chemical industry, says Song, with demand measured in the hundreds of millions of tonnes per year.

Song is also developing palladium–copper catalysts that can efficiently turn CO₂ and hydrogen into methanol. “Mother Nature makes biomass and it takes months and years,” he says. “We can do it in seconds.”

to do better, estimates Aresta. And even then, he says, the conversion will probably not have a net benefit for climate unless the manufacturing plants can make extensive use of cheap energy from solar, wind and other renewables. “If we don’t use these kinds of energy, we will never be able to go to large-scale deployment,” he says.

FUELS AND FUTURE

Some companies are already trying to do just that. In Reykjavik, for example, Carbon Recycling International (CRI) makes methanol from CO₂ and hydrogen using the emissions and electricity from a neighbouring geothermal power plant. The CO₂ in this case comes not from fossil fuel but from carbonate rocks baking in Earth’s heat deep underground. The hydrogen comes from the electrolysis of water — a source that would be prohibitively expensive if the power plant’s electricity were not so abundant. CRI currently produces about 5 million litres of methanol per year, or about 1.5% of the global production of methanol,

which has 95% less CO₂ emissions than petrol. The advantage for Iceland is that methanol, being a liquid fuel, is a much easier way to export the country’s wealth of geothermal energy than, say, laying an underground electricity cable to Europe. But for CRI, the plant is a test bed for wider deployment of its CO₂-to-fuels technology, which it plans to market in other countries such as Germany as a way to capture and store large amounts of renewable energy at times of low demand. “It’s a matter of accessing energy at the right time,” says K.-C. Tran, cofounder and chief executive of CRI.

In most places, because renewable energy isn’t widely and cheaply available, the fuels from CO₂ are still not competitive with those currently being made from crude oil, says Jim Yang Lee, a chemical engineer at the National University of Singapore. “It solves certain, localized problems, assuming you have certain resources,” he says, “but it’s not solving the global CO₂ problem.”

The response from proponents is that it does not have to solve the problem all on its own: if some researchers manage to make progress with CO₂-to-fuels conversion, they argue, while CO₂-based chemical manufacturing continues to expand, CO₂ utilization could make a noticeable dent in greenhouse-gas emissions in the near future.

This year, Styring evaluated various scenarios of CO₂ utilization³. If 100% of urea, 30% of minerals, 20% of specific chemicals and polymers, 10% of methane and 5% of diesel and aviation fuels were supplied by currently known CO₂-utilization methods, he estimates that around 1.34 gigatonnes of CO₂ would be consumed per year. This equals around 83% of the IPCC’s 2030 global target for emissions reductions through CO₂ capture and storage. “These are very conservative estimates,” says Styring. “It is likely that the impact will be much greater.”

For this to happen, researchers acknowledge that it will take far more than clever chemistry. Governments and industries worldwide might have to consider CO₂ utilization along with storage, or even implement a carbon tax to help synthetic fuels stay competitive. “If you have to pay for CO₂ to go into the atmosphere, then the story will change,” says Song.

Aresta is optimistic. “If we are able, in 20 years from now, to use solar and wind energy in a concrete way,” he says, “and use it to bring back CO₂ into fuels, we can very, very easily reduce the emissions of CO₂ by greater than 10%.” ■

XiaoZhi Lim is a freelance writer in Singapore.

1. Aresta, M., Dibenedetto, A. & Angelini, A. J. *CO₂ Util.* **3–4**, 65–73 (2013).
2. Sattthawong, R., Koizumi, N., Song, C. & Prasassarakich, P. *Catal. Today* **251**, 34–40 (2015).
3. Armstrong, K. & Styring, P. *Front. Energy Res.* **3**, 8 (2015).

COMMENT

UNIVERSITIES A call for better pedagogy in Muslim nations **p.634**

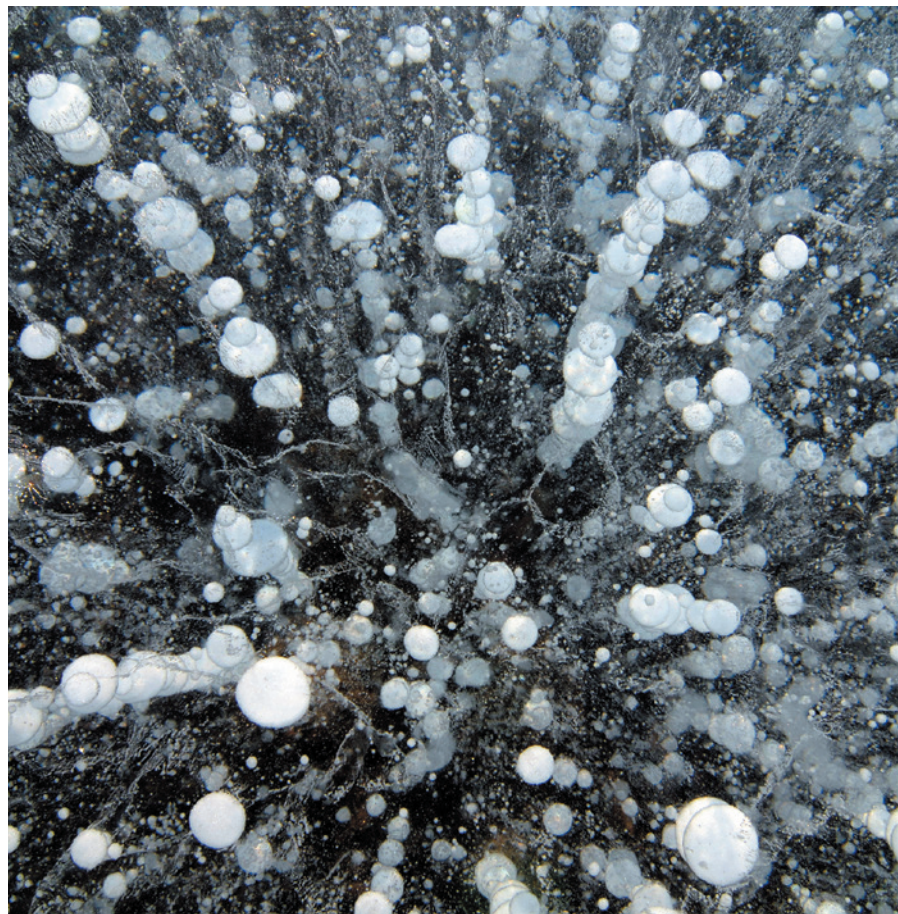


PHOTOS Wartime at Paris's natural history museum captured by Doisneau **p.637**

EPISTEMOLOGY Guide to uncertainty from doyenne of European science policy **p.638**

BADGERS Will upcoming culls be too small to tackle bovine tuberculosis? **p.640**

CLIFF LEIGHT/GETTY



Even extreme environments such as Antarctic ice lakes host microbes.

Create a global microbiome effort

Understanding how microbes affect health and the biosphere requires an international initiative, argue **Nicole Dubilier, Margaret McFall-Ngai and Liping Zhao.**

Microbes have been discovered on Earth wherever anyone has looked for them, from the boiling waters of Yellowstone's hot springs in Wyoming to the depths of cold, dark Antarctic lakes

under 800 metres of ice. A holistic understanding of the role of Earth's microbial community and its genome — its microbiome — in the biosphere and in human health is key to meeting many of the challenges that

face humanity in the twenty-first century, from energy to infection to agriculture.

Recognizing this, a group of leading US scientists this week proposes¹ the creation of a Unified Microbiome Initiative (UMI). The UMI would bring together researchers and representatives from public and private agencies and foundations to study the activities of Earth's microbial ecosystems.

The UMI is conceived as a US initiative; springing from meetings sponsored by the White House Office of Science and Technology Policy and the Kavli Foundation of Oxnard, California. But Earth's biome is not defined by national borders, and efforts to unlock its secrets should go global.

We believe that to be successful, microbiome research will require a coordinated effort across the international community of biologists, chemists, geologists, mathematicians, physicists, computer scientists and clinical experts. As three scientists working in three countries — Germany, China and the United States — we call for an International Microbiome Initiative (IMI) supported by funding agencies and foundations around the world, in addition to the UMI. This would ensure the sharing of standards across borders and disciplines, and bring cohesion to the multitude of microbiome initiatives that exist.

MICROBIAL REVOLUTION

Science is only just realizing the full importance of the microbial world. This is thanks to developments such as low-cost high-throughput sequencing; advances in sample preparation that allow researchers to sequence genomes from individual cells as well as from microbial communities; improvements in computing power and imaging technologies; and the development of bioinformatics tools to help make sense of the data.

Thus biologists are gaining insight into the identity and function of microbes that cannot be grown in the laboratory — the vast majority of Earth's microbiome. Currently only 35 bacterial and archaeal phyla are recognized on the basis of classical approaches to microbial taxonomy. Sequencing efforts in the past few years have pushed the number closer to 1,000 (ref. 2).

Newfound groups of bacteria are throwing old assumptions about the tree of life into question, and revealing vast holes in our



► understanding of the planet's biosphere and its evolution. The discovery in 2003 of giant viruses with hundreds or even thousands of genes shattered the existing definition of living organisms³. (Viruses had long been considered to straddle the line between living and non-living things because of their extreme reliance on host genes.)

It is also becoming clear that microbes provide ecosystem services that are crucial to local and global sustainability. The microbiota in and on crops, trees and other plants, and in the soils in which these grow, provide nitrogen, phosphorus and other essential nutrients. They break down pollutants and suppress the activity of pathogenic microbes. Recognizing the untapped power of soil and plant microbiomes in enhancing agricultural productivity, companies such as Monsanto are investing millions of dollars in research and development in this area.

Microbes in the oceans produce 50% of the oxygen we breathe, and — through photosynthesis — remove roughly the same proportion of carbon dioxide from the atmosphere. They also remove up to 90% of methane from the world's oceans. Over the past decade, research cruises such as Tara Oceans and the Global Ocean Sampling Expedition have sampled, sequenced and analysed the ocean's microorganisms. These have provided insight into the roles that marine bacteria, archaea, viruses and eukaryotic microbes have as global primary producers that provide nutrition at the base of the food chain; remineralization (the transformation of organic molecules into inorganic forms); and the deposition of carbon on the sea floor.

Some of the most profound insights in the

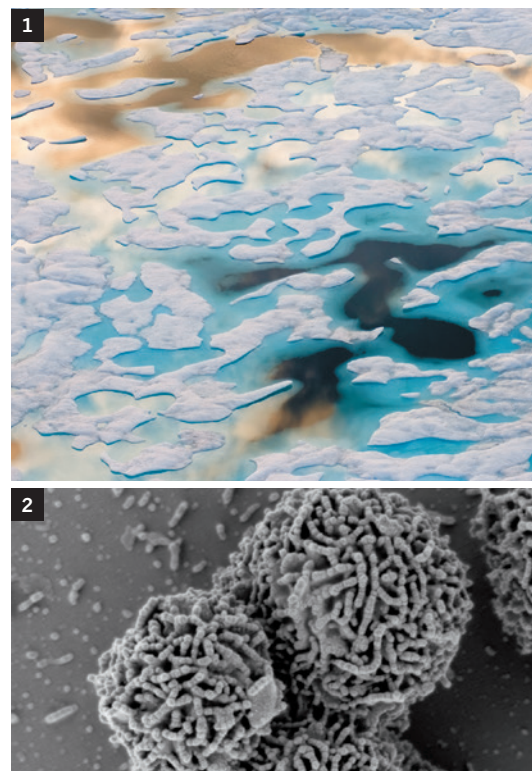
crucial role of microbes for human well-being have emerged from analyses of the microbes on and in our bodies — their genomes, transcriptomes, proteomes and metabolomes. (These are analyses of genes, RNA molecules, proteins and chemical metabolites). Complex gut communities, for instance, protect us from disease, provide nutrition, and affect our development even before birth⁴.

STUMBLING BLOCKS

There are two major stumbling blocks to advancing our understanding of microbes' role in the biosphere. First is the fragmentation of the life-sciences field. Second is a lack of coordination among the various microbiome research endeavours under way.

Disciplinary silos are problematic because any attempt to understand anything about plants or animals needs to be rooted in microbiology. For example, for decades, circadian rhythms in the mammalian gastrointestinal system were studied in the context of human physiology and gene expression. Yet in the past two to three years, biologists have discovered that daily cycles in the motility of the gut — the production of digestive enzymes, gene expression in gut cells and so on — rely on the activities of gut microbiota. Compounds produced by microbes either cause changes in the gut directly or pass into the host's bloodstream and influence the central nervous system, possibly through neural, hormonal and immune pathways⁵.

The first hint of this came from the discovery (led by one of us, M.M.-N.) that bioluminescence and other products of the marine luminous bacterium *Vibrio fischeri* regulate the expression of a circadian



cryptochrome gene in squid⁶. This highlights the value of investigating host-microbe relationships from all branches of the tree of life, including those in which only a single symbiotic species is involved.

PROJECT PROBLEMS

Since a 2005 workshop in Paris, at least eight programmes have been established to study the human microbiome. These include the US Human Microbiome Project, the Canadian Microbiome Initiative, MetaHIT (involving the European Union and China) and the Human Metagenome Consortium in Japan.

These initiatives have generated vast amounts of data that are not easily comparable. For example, many studies on human microbiota identify species (or operational taxonomic units) and map evolutionary relationships using the 16S ribosomal RNA gene. Differences between the primers (which provide a starting point for the DNA synthesis) used to amplify this gene can have a big effect on the sequence data ultimately obtained. And estimates of species numbers can vary by two to three orders of magnitude, even when applied to the same sample, because different software packages are being used to analyse the amplified genes⁷.

This lack of consistency in approaches means that effective comparisons and interpretations of human microbiota studies are often not possible. The International Human Microbiome Consortium, established in 2008, and the International Human Microbiome Standards project, launched in 2011, have attempted to address some of these

GOING GLOBAL

Four functions of an International Microbiome Initiative

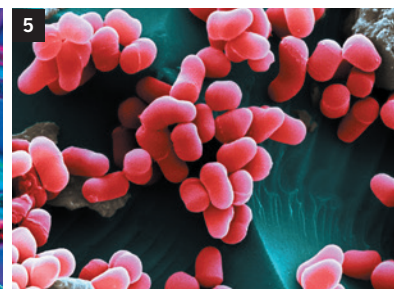
Guidelines. Create a working group to oversee the development and implementation of guidelines for the study of microbiomes, drawing on and developing those already established by other initiatives such as the Earth Microbiome Project. The group would set standards for methods, data analysis, data sharing and intellectual-property rights, and would partner with funding agencies and publishers to ensure that researchers follow the agreed guidelines.

Priorities. Develop a common research agenda, with the goal of enabling comparative analyses that range from local to global scales. For instance, one priority could be to increase the number and diversity of people sampled in studies of the human microbiome.

Tools. Identify new cross-disciplinary methods for microbiome studies. Examples include imaging techniques ranging from confocal advances to cryotomography — which can resolve subcellular structure and reveal microbial cell function — and ways to monitor the production and exchange of microbial metabolites.

Forums. Establish platforms for the discussion and exchange of research within and between nations, the development of programmes for training the next generation of microbiome scientists, and the establishment of outreach projects to educate and engage the general public. A good model is the Ocean Sampling Day's citizen's science campaign, which recruits citizens to help obtain environmental data.

Microbes such as *Planococcus halocryophilus* (2), cyanobacteria (4) and *Arthrobacter crystallopoietes* (5) are found everywhere from soils (6) to the Arctic's Chukchi Sea (1) to Yellowstone's Grand Prismatic Spring (3).



issues. But, especially in the case of the standards project, many national research projects were already well under way. Both initiatives are struggling with issues about data sharing and differences between national policies in ownership and property rights.

GLOBAL SOLUTION

We think that an IMI could do a better job. The study of any microbiome demands myriad collaborations. These should involve basic and applied biologists, including those with expertise in microorganisms or in higher organisms; informaticians and mathematicians, who can develop methods that extract information from the mountains of sequence data; and chemists, physicists and engineers. Physical scientists are needed to find new ways to measure and manipulate the compounds that microbes produce and exchange with their biotic and abiotic surroundings.

An IMI would be pivotal in bringing together all these experts and allow scientists to move beyond cataloguing, which has characterized much of microbiome work so far. And, it would be limiting to recruit such a vast range of intellectual, scientific and technological expertise from one country alone.

IMI participants could use comparative approaches to reveal the factors that underlie the structure and function of microbiomes on local to global scales. An example of the potential of comparative approaches comes from studies of Native Americans⁸ and African hunter-gatherers⁹. It seems that these groups have a much higher diversity

of microbial partners (a known correlate with better health) than do people living in industrialized societies.

By pooling data from scientists from around the world, an IMI would generate much more knowledge than could one country alone. Thanks to the falling costs of sequencing machines, individual labs will probably soon produce more data than the conventional large sequencing centres, such as the Joint Genome Institute in Walnut Creek, California. Yet for any one laboratory, sample sizes might be restricted, and researchers might have only limited bioinformatics capacity.

An IMI could encourage the integration of data across institutions and nations. This is especially important for countries that may not have the funds to invest in their own global-scale projects. For example, cloud-computing platforms would allow people to upload and analyse sequencing data as soon as they are available. The IMI could also control and organize access to metadata (the associated host disease phenotype data, for a human gut microbiota sample, for instance) without which meaningful interpretation of the data is not possible. This could also be a way to safeguard the intellectual property of researchers, funding bodies and nations.

Most importantly, an IMI is essential when it comes to solving problems that affect

the biosphere. Although processes involving microbiomes vary from place to place, the impact of such processes can often be felt globally. Potent greenhouse gases, such as nitrogen oxides produced by denitrifying bacteria in overfertilized Chinese farming lands or methane released by archaea in the millions of ruminant animals in Australia and New Zealand, may have contributed substantially to global warming. Billions of tonnes of human-made toxic chemicals have overwhelmed the degrading and recycling capacity of microbiomes. And the imprudent use of antibiotics has contributed worldwide to epidemics of chronic diseases, such as obesity, diabetes and cancer. The solutions to some of these problems may come out of local research, but an IMI is essential for ensuring that comparable data are produced from efforts throughout the international scientific community.

We do not have all the answers when it comes to designing an IMI. In fact, we think that the first step to launching such a project should be the bringing together of leading microbiome researchers from across the globe to discuss its goals. However, even at this stage, some elements seem crucial (see 'Four functions of an International Microbiome Initiative').

As long as communities are prepared to start afresh and conform to new ground rules, we believe that an IMI could succeed where other efforts to achieve standardization and coordination have struggled. Such a project would be ahead of the curve, because the tools needed to explore the world's microbes

"We urge scientists to help make an IMI happen by sharing their data."

have only recently become available. Also, if multiple working groups with representatives from across the life sciences were established — similar to those set up by an effort to assess marine organisms called the Census of Marine Life — an IMI could be much broader in scope than any pre-existing programme.

Finally, an IMI would be able to solve the data sharing and intellectual-property issues that have been stumbling blocks for previous efforts, by organizing and controlling access to the metadata that are so essential for interpreting results, publishing papers and filing patents.

It is crucial that the IMI is launched quickly to avoid corrective actions having to be applied after-the-fact to national efforts. We invite private foundations that have been pivotal in mounting international research efforts to support an IMI. These include the Gordon and Betty Moore Foundation with its Marine Microbiology Initiative, the Alfred P. Sloan Foundation with its Census of Marine Life, the Kavli Foundation with its Brain Initiative, and the Bill & Melinda Gates Foundation with its Global Health Program.

We also encourage national funding agencies to open up their programmes to international collaborations and to adopt any standards established by an IMI. Finally, we urge scientists to help make an IMI happen by sharing their data.

So much can be gained by creating an IMI. Further uncoordinated national microbiome programmes will almost certainly waste research efforts and taxpayers' money. Let's transcend national silos and gain universal insights that will benefit all humankind. ■

Nicole Dubilier is a director at the Max Planck Institute for Marine Microbiology in Bremen, Germany. **Margaret McFall-Ngai** is professor and director of the Pacific Biosciences Research Center at the University of Hawaii at Manoa, Honolulu, Hawaii, USA. **Liping Zhao** is professor of microbiology at Shanghai Jiao Tong University, Shanghai, China. e-mail: ndubilie@mpi-bremen.de

1. Alivisatos, A. P. *et al. Science* **350**, 503–504 (2015).
2. Yarza, P. *et al. Nature Rev. Microbiol.* **12**, 635–645 (2014).
3. La Scola, B. *et al. Science* **299**, 2033 (2003).
4. Aagaard, K. *et al. Sci. Transl. Med.* **237**, 237ra65 (2014).
5. Mukherji, A., Kobiita, A., Ye, T. & Chambon, P. *Cell* **153**, 812–827 (2013).
6. Heath-Heckman, E. A. C. *et al. mBio* **4**, e00167-13 (2013).
7. Chen, W., Zhang, C. K., Cheng, Y., Zhang, S. & Zhao, H. *PLoS ONE* **8**, e70837 (2013).
8. Clemente, J. C. *et al. Sci. Adv.* **1**, e1500183 (2015).
9. Schnorr, S. L. *et al. Nature Commun.* **5**, 3654 (2014).



The world's oldest continually operational university was founded in Fes, Morocco, in AD 859.

Revive universities of the Muslim world

To boost science, higher-education institutes must give students a broad education and become meritocratic, say **Nidhal Guessoum** and **Athar Osama**.

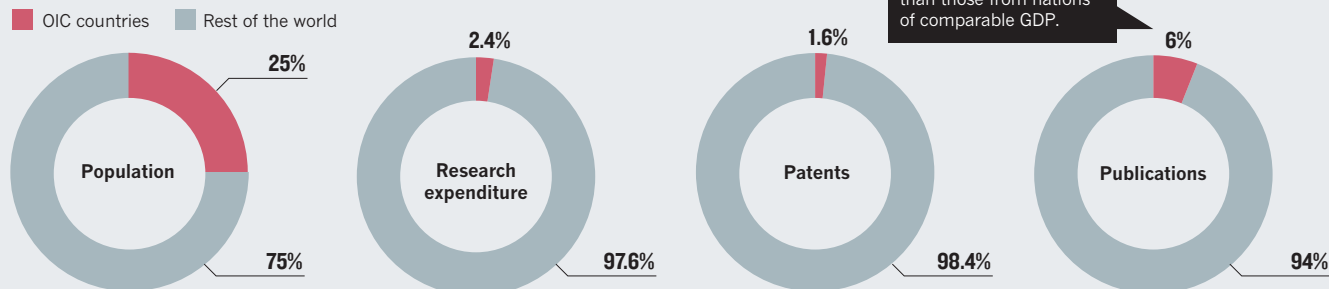
The Islamic civilization lays claim to the world's oldest continually operational university. The University of Qarawiyyin was founded in Fes, Morocco, in AD 859, at the beginning of an Islamic Golden Age. Despite such auspicious beginnings, universities in the region are now in dire straits, as demonstrated by a report we have authored,

released this week (see go.nature.com/korli3).

The 57 countries of the Muslim world — those with a Muslim-majority population, and part of the Organisation of Islamic Cooperation (OIC) — are home to nearly 25% of the world's people. But as of 2012, they had contributed only 1.6% of the world's patents, 6% of its academic publications, and 2.4% of

QUARTER DECK

Investing on average 0.5% of their gross domestic product (GDP) on research and development (less than one-third of the global average), the Organisation of Islamic Cooperation (OIC) countries' science output lags, as measured by patents and publications.



the global research expenditure^{1,2} (see 'Quarter deck'). There have been only three Nobel laureates in the sciences from OIC countries; today these nations host fewer than a dozen universities in the top 400 of the many world rankings, and none in the top 100.

To assess this situation, for the past two years we have led an international non-governmental and non-partisan task force of experts, formed by the Muslim World Science Initiative. The task force was chaired by Zakri Abdul Hamid, science adviser to the prime minister of Malaysia. It included a dozen experts and scholars — including policymakers, vice-chancellors, professors, and science communicators — from around the world.

Our work confirmed many widely known problems, as highlighted by reports such as the Royal Society's 2014 *Atlas of Islamic World Science and Innovation*². For example, OIC countries on average invest less than 0.5% of their gross domestic product (GDP) on research and development (R&D). Only Malaysia spends slightly more than 1% (the world average is 1.78%; most advanced countries spend 2–3%). Students in the Muslim world who participate in standardized international science tests lag well behind their peers worldwide, and the situation seems to be worsening^{3,4}.

Our report highlights an even more problematic situation. University science programmes are using narrow content and outdated teaching methods. In most OIC countries, students are channelled into science or non-science streams around the age of 14, and their education thereafter is completely binary: science and technology students receive little in the way of humanities, social-science, language or arts education, and vice versa. Only one university in the region offers a programme in 'science and technology studies': the University of Malaya in Kuala Lumpur.

To become beacons in society, OIC universities need to revitalize their teaching methods and meld science with liberal arts such as history and philosophy. For universities to

become truly meritocratic, they must develop new ways of assessing faculty members to reward valuable research, teaching and outreach. And for this to happen, governments must give universities more autonomy.

QUANTITY AND QUALITY

Our task force gathered data on science production for the 20 OIC countries that together have represented more than 90% of OIC scientific productivity over the past two decades. From the period 1996–2005 to 2006–15, most countries doubled or tripled their production of science papers. Qatar's output rocketed by a factor of 7.7, and Iran's by 7.6. But the number of scientific papers produced remains below the average of countries with similar GDP per capita. We found an average of 4.2 papers per dollar of GDP per capita for our OIC sample in the most recent decade, compared to an average of 8.6 for a group of 4 peer countries such as Brazil, Spain, South Korea, South Africa and Israel (see Supplementary Information; go.nature.com/4o3itm).

Papers from these OIC countries are cited less frequently than those from other nations. The average was 5.7 citations per paper for 2006–15, compared with 9.7 for South Africa and 13.8 for Israel, countries with a comparable GDP per capita. A list of the 100 most-cited papers since 1900 has none with a lead author from a Muslim-majority nation (see *Nature* **514**, 550–553; 2014).

Scientific research must be relevant and responsive to society's intellectual and practical needs. This dual goal seems to be out of sight — and often out of consideration — for most academic institutions in the region.

For scientists and engineers to be creative, innovative and able to engage with questions of ethics, religion and the wider social purpose of research, students must receive a broad, liberal-arts-style education⁵. A few

institutions attempt to relate their students' learning to their cultural backgrounds and contemporary knowledge. In the early 1970s, Tehran's Sharif University of Technology began a rich programme melding Islamic history, philosophy and culture with science and engineering. Its graduate programme in the philosophy of science remains the only one in the OIC that we are aware of. It is perhaps no coincidence that the most recent Times Higher Education world university rankings named Sharif University as the top Iranian university and number eight in the OIC.

In recent years, US-style liberal-arts establishments have been set up in the region, modelled on the long-running and respected American University of Beirut and the American University in Cairo. One such is the American University of Sharjah (AUS) in the United Arab Emirates (UAE), which this year ranked seventh in the QS Rankings of universities in the 22 Arab countries. Fully home-grown and self-funded and with no formal affiliation with a US institution, the AUS requires science and engineering students to take roughly one-third of their required 40 or so courses in humanities, social sciences, language and communication.

Habib University, founded last year in Karachi, Pakistan, also follows this model. Here, science and engineering students must take courses such as 'Understanding Modernity' and 'Hikma 1 & 2' — a two-course sequence that translates as 'traditional wisdom' — as well as many others that seek to create rounded rather than narrow engineering and science professionals. Other educational establishments should follow suit.

CURRICULAR INNOVATION

Science classes themselves have serious problems. The textbooks used in OIC universities are often imported from the United States or Europe. Although the content is of a high standard, they assume a Western experience and use English or French as the language of instruction. This disadvantages many students, and creates a disconnect between their

"Science classes themselves have serious problems."



Science students at the American University of Sharjah must take humanities and social-science courses.

education and culture. To encourage the production of higher-quality, local textbooks and other academic material, universities need to reward staff for producing these at least as much as they do for research publication.

Some basic facts are seen as controversial, and marginalized. Evolution, for example, is usually taught only to biology students, often as “a theory”, and is rarely connected to the rest of the body of knowledge. One ongoing study has found, for example, that most Malaysian physicians and medical students reject evolution (see go.nature.com/38cswo). Evolution needs to be taught widely and shown to be compatible with Islam and its culture⁶. Teaching the philosophy and history of science would help, too.

The global consensus is that enquiry-based science education fosters the deepest understanding of scientific concepts and laws. But in most OIC universities, lecture-based teaching still prevails. Exceptions are rare. One is the Petroleum Institute, an engineering university in Abu Dhabi, UAE, where the faculty has created a hands-on experience with positive results on student interest and enrolment, particularly of women.

Another problem is that faculty members rarely — if ever — receive any training or evaluation in pedagogy. This is true elsewhere in the world, but change is harder in many OIC nations. In most, curriculum changes, faculty appointments and promotions are set by ministry rules and decided by centralized commissions and bureaucracies. This leaves little room for universities to innovate.

THE WAY FORWARD

Universities in OIC nations need to be granted more autonomy to transform themselves into meritocracies that strive

for scientific excellence and then lead rather than follow the winds of change towards greater transparency and meritocracy within their societies.

Universities need to promote the right metrics, so that they do not inadvertently encourage plagiarism and junk science through pressure to publish. The region needs consistent data on science student and faculty profiles, curricula, pedagogy, language of instruction and so on, akin to what the Institute of Statistics of the United Nations Educational, Scientific and Cultural Organization collects — but at a fine-grained, university level. This is a task that must be undertaken by national or transnational bodies, such as the Islamic World Academy of Sciences (IAS) or the Islamic Educational, Scientific and Cultural Organization (ISESCO).

We also call for reform of science curricula and pedagogy. Universities need to deliver more multi-disciplinary, exploratory science education. A good start would be training for university teachers, with workshops on new tools and approaches. Barriers need to be broken between departments and colleges and new programmes constructed. Professors need to be free to teach topics that are not tightly regulated by ministries.

There are grassroots efforts across the Muslim world to stimulate curiosity about science among students of all ages, operating without much government support. Ahmed Djebbar, an emeritus science historian at the University of Lille in France, has

“Professors need to be free to teach topics that are not tightly regulated by ministries.”

constructed an online, pre-university-level course called ‘The Discoveries in Islamic Countries’ available in three languages⁷, which relates science concepts to great discoveries and stories from the Islamic Golden Age. Such courses should be scaled up and shared by many institutions.

Universities will need to implement reforms individually. We hope that the inspiration from a few islands of excellence will, in time, turn the tide of public and political opinion. There is precedent. In Pakistan, two private universities established in the 1980s — the Aga Khan University and Hospital in Karachi and Lahore University of Management Sciences — revolutionized medical and business education within a decade of their creation. Students elsewhere began demanding the standard set by these educational pioneers. The same can be done for science.

Our task force is putting out an open call for universities across the Muslim world to join a voluntary Network of Excellence of Universities for Science (NEXUS), to be launched early next year. This peer group will be managed by the task force and housed in science adviser Zakri’s office. We plan for NEXUS to run summer schools for university administrators, to monitor the progress of reforms at participating universities, and to issue a peer report card that will assess the performance of the universities in meeting milestones, thus recognizing and inspiring further improvements. True transformation will require much broader action from ministries, regulators and funding agencies, and these may be the most resistant to change.

Without tough reforms, the dream of a scientific revival in the Muslim world will remain just that. ■

Nidhal Guessoum is professor of physics and astronomy at the American University of Sharjah, United Arab Emirates.

Athar Osama is an honorary senior associate at the UCL Institute of Education, London, UK, and the founder of the Muslim World Science Initiative.
e-mails: nguessoum@aus.edu; athar.osama@gmail.com

1. Royal Society. *A New Golden Age? The Prospects for Science and Innovation in the Islamic World* (Royal Society, 2010).
2. Royal Society. *The Atlas of Islamic World Science and Innovation* (Royal Society, 2014).
3. Martin, M. O., Mullis, I. V. S., Foy, P. & Stanco, G. M. *TIMSS 2011 International Results in Science* (TIMSS & PIRLS International Study Center, 2012).
4. OECD. *PISA 2012 Results in Focus* (OECD, 2014).
5. Bloom, D. E. & Rozovsky, H. *Liberal Educ.* **89**, 16–23 (2003).
6. Guessoum, N. *Islam’s Quantum Question: Reconciling Muslim Tradition and Modern Science* (IB Tauris, 2011).
7. Djebbar, A. *The Discoveries in Islamic Countries* (Éditions le Pommier, 2009).



The Funeral Procession of the Jaguar: a technician transports a specimen in 1943.

PHOTOGRAPHY

Sedition in the stores

Laura Spinney extols Robert Doisneau's haunting images of the Paris natural history museum under occupation.

In 1942, French photographer Robert Doisneau (perhaps best known for his image of a couple kissing outside the Hotel de Ville) was commissioned to record life behind the scenes at the various arms of the National Museum of Natural History (MNHN) in Paris. Most of the images have never been published. They are a unique document of the work of a research institute in occupied France during the Second World War. Now, a small

jewel of an exhibition brings them out of the stores where they were taken, and places them in the limelight where they belong.

Doisneau was a member of the French Resistance; he produced fake identity papers for his comrades-in-arms. The commission to photograph two Paris museums, the botanical gardens and a zoo that are part of the MNHN, was offered to him by the influential publisher

Robert Doisneau:
A Photographer at the Museum
French National
Museum of Natural
History, Paris.
Until 18 January 2016.

Maximilien Vox (real name, Samuel Monod), acting on behalf of the Vichy government. Sympathetic to the Germans, this puppet regime wanted to

vaunt the vitality of French intellectual life under its beneficent new rulers.

Why did Doisneau agree to such a dubious assignment? Recently returned from the army, he probably just needed the cash. His first baby had only recently been born, and a commission from Vox was not something that a young photographer turned down. Furthermore, lauding France's academic excellence need not have struck him as a betrayal.

What Doisneau found as he toured the museums and gardens was a vibrant research institute — despite, rather than because of, the intrusion of world events. Paul Rivet, director of the MNHN's Museum of Man, was in exile in Colombia. Others had just returned from military service or prisoner-of-war camps — including the palaeontologist Camille Arambourg, now remembered for defending Neanderthals against accusations of simian brutishness.

A demobilized botanist, André Guillaumin, was searching for coal to heat the vast greenhouses. A major effort was under way to reorganize the collections, which were just starting to return, having been evacuated in 1939. Publication had been slowed but not stopped by the censors. At the Museum of Man, a resistance cell had been dismantled and its members executed or deported.

Emerging from such tensions, the images take on extra significance. Doisneau wrote later that he was struck by the contrast between the moment of history he inhabited — of which his growling stomach served as a constant reminder — and the geological epochs spanned by the collections. He used that contrast to powerful effect.

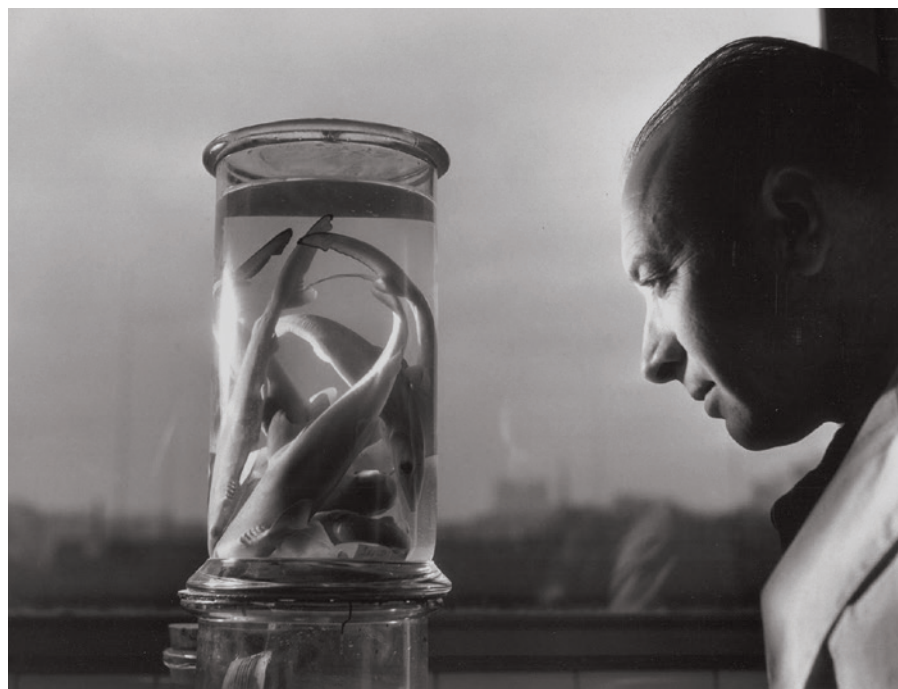
Some of the images seem downright insolent, such as that of Paul Budker of the Laboratory of Fish and Colonial Animals gazing into a jar of baby sharks — a Frenchman inspecting imprisoned predators. Others find a wistful wisdom in scenes from the museum's daily life. One such is *The Funeral Procession of the Jaguar*: the beast is pushed in a wheelbarrow over cobbles to the taxidermy department. Another, showing a white-coated woman with a wizened corpse on her hip and a faraway look in her eyes, Doisneau entitled *The Surprising Lightness of a Peruvian Mummy*.

Vox had envisaged a collection ▶

► called *The Face of Science*. Overtaken by events, it never saw the light of day. In November 1942, Allied forces landed in North Africa, prompting the Germans to invade previously unoccupied southern France, rendering the Vichy government impotent. The photos were consigned to the museum library.

In 1990, the museum invited Doisneau back to complete his project. This postscript was a good idea: the contrast between the two sets of photos speaks volumes. Doisneau was in his late seventies and famous. The museum, too, had changed, and Doisneau delighted in discovering its three new subterranean floors of storage. The later images are as closely observed as the earlier ones. But now — as in a picture of a stuffed gorilla in a lift, emerging from or descending into the museum's bowels — the irony is less loaded, and the delight floats free. ■

Laura Spinney is a writer and science journalist based in Paris.
e-mail: lfspinney@gmail.com



ATELIER ROBERT DOISNEAU

Paul Budker of the Laboratory of Fish and Colonial Animals with a jar of baby sharks in 1943.

SCIENTIFIC METHOD

Tales of the unexpected

Stuart Firestein relishes Helga Nowotny's study of uncertainty in science and society.

For scientists, uncertainty is a norm. Experiments begin with uncertainty (why else do them?), and even when they are 'successful', the results contain only a range of certainty and a range of confidence about that certainty. Yet in the world outside the laboratory, uncertainty is perceived as negative — not a data point, but a failing, effectively no better than not knowing.

Social scientist Helga Nowotny, former president of the European Research Council, has written *The Cunning of Uncertainty*, a tour of the phenomenon and its value — to the individual researcher, to the infrastructure of research and to society. Coping with uncertainty, Nowotny declares, must become a collective achievement. Otherwise, the scientific community risks becoming an elite, with all the suspicion and mistrust that that engenders.

As Nowotny shows in numerous examples from the social sciences, historical literature and current media, society's misunderstanding of uncertainty has already led to confusion, distortion and politicization of science (in the debates on tobacco and on climate change, for instance). Billions of dollars of public and

private money are poured into research, largely on the false assumption that science provides cold, hard, immutable facts. This attitude is reinforced by an educational system that treats science as an immense 'fact tract' to be memorized (and then largely forgotten), producing a populace that believes science is about answers, rather than questions.

Nowotny deserves high praise for bringing a discussion of the uncertainty around uncertainty into the public arena. She does so with remarkable aplomb given the subject's slipperiness. She challenges scientists to take a leading role in setting straight the distorted views of certainty in science, and then spreading the word to educational and political policymakers.

I was puzzled at first by Nowotny's use of "cunning", a term denoting a worrisome craftiness — cleverness mixed with predatory wiliness. However, Nowotny has nailed

it: uncertainty has a duality. It is a space that allows creativity, but it is fraught with insecurity. Nowotny herself struggles to disentangle uncertainty from the unexpected, the unpredictable, the dangerous. She sometimes fails, but the exercise is instructive. We take too much for granted, she shows, in defining uncertainty; there are many subtleties and layers to it. From antiquity, people used prognosticators and magicians to look into the future; futurologists are still consulted. Yet people insist on 'spoiler alerts' before sports results are announced, or in film or book reviews; and few want to guess, much less know, the exact moment and circumstances of their too-certain deaths. Definitely cunning.

Nowotny examines this tension through historical examples of how policy is set in areas of uncertainty. These are: reproductive technologies such as *in vitro* fertilization; stem-cell science; and personalized medicine. With the exception of the last example, these seem like old stories. A discussion of more current and contentious policies, such as genetically modified organisms or nuclear power, might have been more instructive. However, her analyses of big-data programmes (the trumpeted then debunked



The Cunning of Uncertainty
HELGA NOWOTNY
Polity: 2015.

Google Flu Trends is an excellent example) and nonlinear complex systems (such as finance, energy or pandemics) as gushers of uncertainty and contemporary spaces of innovation make for deep and fruitful reading.

Although she gives us chapter headings like those of a self-help book ('Craving for certainty', 'The odds for tomorrow'), Nowotny refuses to be

"Treating science as a 'fact tract' to be memorized produces a populace that believes science is about answers, rather than questions."

prescriptive — at least in this volume. Rather, she presents us with many ideas and numerous angles to chew over. My personal favourite is her tracking of public attitudes to risk, from viewing it as both negative and positive to seeing it as an almost completely negative factor that must be reduced. Her ideas about positive risk and why it must be increased offer a refreshing perspective in this compliance-oriented world.

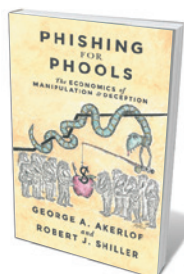
This is, above all, a book of ideas, not a policy manual, even though Nowotny would obviously like to see changes in funding, economic and management policies. She is well-equipped to lead us in this battle; throughout an illustrious academic career, she has served in numerous policymaking positions, for example in the European Science Foundation. And that is perhaps the one thing that I miss in this book: a more personal tour of uncertainty with a traveller who has come up against it in policy, funding, education and research. It is a shame that Nowotny does not occasionally put down the careful scholarly pen and take up the memoirist's.

As a scientist, I am at home with uncertainty. I like that the gathering of knowledge inevitably reveals new and unexpected bits of a vast unknown akin to the dark matter of epistemology. The real cunning of uncertainty lies in how it increases through every attempt to reduce it.

As for certainty, wherever you find it, you can be sure that a demagogue or dictator is nearby. French author and Nobel laureate André Gide advised that we should believe those who seek the truth — and doubt those who claim to have found it. This is one prescription that Nowotny courageously follows. ■

Stuart Firestein is the former chair of the Department of Biological Sciences at Columbia University in New York City. His latest book is *Failure: Why Science Is So Successful*.
e-mail: sjf24@columbia.edu

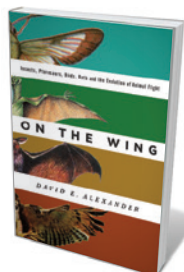
Books in brief



Phishing for Phools: The Economics of Manipulation and Deception

George A. Akerlof and Robert J. Shiller PRINCETON UNIV. PRESS (2015)

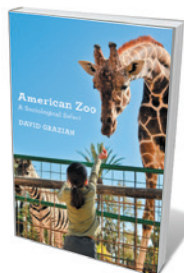
In this acerbic dissection of free-market economics, Nobel-prizewinning economists George Akerlof and Robert Shiller trash the "invisible hand" theory, which claims that self-interest promotes social benefits. They reveal market economies as rife with trickery — "phishing" luring "phools" to make poor choices. The two pool their economic wisdom to analyse arenas from food buying and politics to the financial crisis that has plagued us since 2008. A needed call for sceptical economics and financial mindfulness.



On the Wing: Insects, Pterosaurs, Birds, Bats and the Evolution of Animal Flight

David E. Alexander OXFORD UNIV. PRESS (2015)

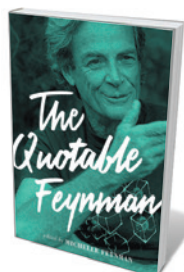
How do dragonflies, vultures or fruit bats fly? Biomechanics specialist David Alexander reveals all in this study of flight and the evolution of wings in pterodactyls, bats, birds and insects. Aloft with Alexander, we learn how pilots have observed phalanxes of swans flying at 8,000 metres, and how the gargantuan pterosaur *Quetzalcoatlus northropi* probably got off the ground. Alexander analyses aerial predation, sex, combat, sleep and even egg-laying; picks at the puzzle of bats' evolutionary relationships; and much more.



American Zoo: A Sociological Safari

David Grazian PRINCETON UNIV. PRESS (2015)

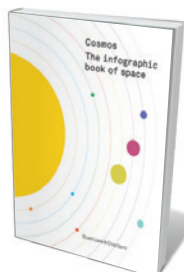
Cultural sociologist David Grazian once studied urban nightlife. Turning to another brand of contained wildness, he immersed himself in zoos. As a volunteer, he clipped a ferret's toenails, bathed tortoises and logged "more working hours of animal husbandry and faecal cleanup" than most professors can boast. His trek through 26 US zoos has yielded a powerful portrait of these conservation-hotspots-cum-living-labs — which end up telling us more about ourselves than about the animals. Peppered with delicious details, such as one zoo's use of the film *Austin Powers* for animal "enrichment".



The Quotable Feynman

Edited by Michelle Feynman PRINCETON UNIV. PRESS (2015)

The Nobel-prizewinning, bongo-playing, exuberant and brilliant physicist Richard Feynman died in 1988. His contributions to science (including the theory of quantum electrodynamics) and science popularization ensure a lasting fame (R. Phillips *Nature* **504**, 30–31; 2013). His daughter Michelle has mined interviews, articles, books and lectures for this collection of quotes on everything from poetry to politics. Feynman's depth and zing leap from the page, as in: "What I am trying to do is bring birth to clarity, which is really a half-assedly thought-out pictorial semi vision thing."



Cosmos: The Infographic Book of Space

Stuart Lowe and Chris North AURUM (2015)

Infographics remain on a roll, offering visual insight into abstruse regions of human knowledge. Astronomers Stuart Lowe and Chris North really lift off in this graphic exploration of all things space. Their depictions of year-by-year spaceflights and space junk shock through sheer numbers, while their takes on interplanetary missions, moons in the Solar System and particularly the polarization of the Milky Way — like a cosmic finger-painting in palest mauve — deliver the facts with aesthetic brio. [Barbara Kiser](#)

Correspondence

Engage the public to stop bear trafficking

Last month, authorities in China confiscated 100 kilograms of bear paws. Many of the animals had been massacred in Russia, then the paws smuggled into China. Neither country is managing to control this monstrous trade.

Bear paws are a coveted delicacy in China, fetching up to ten times the price that they do in Russia. This is despite some species having a second-class protected status under Chinese law, with bear hunting and the sale and eating of bear products strictly prohibited in the country.

The Chinese government must intensify its efforts against wildlife trafficking. Extensive publicity and public education will be the most effective way to kill demand for wildlife delicacies.

Zhengrong Yuan, Yingying Han, Qiang Weng *Beijing Forestry University, China*
qiangweng@bjfu.edu.cn

China Nobel stirs up attack on academies

The first Nobel prize to be awarded to a researcher in China has sparked heated debate in the country (see *Nature* **526**, 174–175; 2015). This centres on the public's long-term dissatisfaction with the Chinese academy system, which consistently failed to recognize the scientific talent and originality of the prizewinner Youyou Tu.

Chinese news reports about Tu's award on 5 October drew attention to her 'three-no' status (that is, no overseas experience, no doctorate, no admission to a national academy — a status known as *yuanshi*). Citizens are indignant that her *yuanshi* application was repeatedly declined, despite her recommendation by the minister of health and her acclaim by international scientific bodies.

The unfairness of the *yuanshi* selection process is widely blamed on cronyism, nepotism and

excessive bureaucracy (see, for example, C. Cao *et al. Science* **341**, 460–462; 2013). China's academic institutions need to reform their moribund practices rapidly or they are likely to overlook other striking individuals and key grass-roots research in the future.
Xin Miao *Harbin Institute of Technology, Harbin, China*
xin.miao@aliyun.com

Badger-cull targets unlikely to reduce TB

Two months ago, the government-advice body Natural England approved further licensed badger culls in parts of the United Kingdom in 2015. The aim is to reduce local badger densities by at least 70% to prevent the spread of tuberculosis (TB) to cattle (see go.nature.com/iutvj). On the basis of the government's badger-population estimates, we calculate that these culls are unlikely to achieve the necessary reduction.

The latest minimum cull numbers derive from the lower 95% confidence bounds on population size estimates. For example, licensees in Dorset are required to kill at least 615 badgers in a population estimated at 879–1,547 animals (95% confidence interval). Killing this number would give an estimated population reduction of between 39.8% and 70% (95% confidence interval).

Equivalent confidence intervals for the 2015 (third annual) Somerset and Gloucestershire culls are, respectively, 50.8–70% and 54–70% relative to the baseline population estimates. It is therefore unlikely that a 70% or greater reduction can be attained by these minimum cull numbers, assuming that the population estimates are accurate.

Evidence from a randomized controlled trial shows that better prospects for the control of cattle TB are offered by badger populations that are either reduced by more than 70% or left undisturbed — and potentially

vaccinated (C. A. Donnelly *et al. Nature* **439**, 843–846; 2006). The choice depends on a range of epidemiological, economic, social and ecological factors.

Christl A. Donnelly *Imperial College London, UK*

Rosie Woodroffe *Institute of Zoology, London, UK*
c.donnelly@imperial.ac.uk
Competing financial interests declared; see go.nature.com/so3gvl.

EU report advises on contentious research

The United States has a de facto moratorium on genetic gain-of-function experiments that could increase the transmissibility or pathogenicity of potentially pandemic agents such as the H5N1 avian influenza virus. In Europe, opinion among scientists is divided on the benefits and risks of such research for policymakers. A new report on these differences by the European Academies Science Advisory Council (EASAC; see go.nature.com/jcdy2w) will help to inform scientists and the public on this globally controversial research.

As well as risk–benefit assessment, the report addresses concerns such as scientific responsibility, research review and management systems, options for national and international biosafety and biosecurity advisory bodies, and the publication of sensitive information. It also highlights areas in which European Union regulations and best-practice guidelines need further consideration, notably those affecting the publication and export of research findings. We recommend an integrated approach to biorisk assessment and management, with responsibilities and action shared among researchers, institutions and funders.

Robin Fears, Volker ter Meulen *EASAC, German National Academy of Sciences Leopoldina, Halle, Germany*
volker.termeulen@mail.uni-wuerzburg.de

Make raw emissions data public in China

China's carbon emissions need to be estimated more accurately if the country is to meet its climate targets and participate in a nationwide emissions-trading scheme after 2017 (see F. Teng *Nature* **525**, 455; 2015). Accounting errors will prevail until the nation's data sources are made transparent, measurable and verifiable.

For example, Teng points to discrepancies between our estimations of coal emissions (Z. Liu *et al. Nature* **524**, 335–338; 2015) and those calculated by the government's data agencies — from information that is not publicly available. The comparison is invalid, however. Our estimates are derived from energy consumption and measured emission factors for coal acquired from 4,845 mines, rather than from final energy consumption and estimated emission factors averaged over all coal types.

Emissions estimates are even more uncertain in economies such as India, Indonesia and South Africa. China's latest investment of 20 billion yuan (US\$3.1 billion) for climate-change mitigation and adaptation in developing countries could help these nations to compile measurable, accurate and accessible carbon emissions data.

Dabo Guan *University of East Anglia, Norwich, UK*
Zhu Liu *Harvard University, Cambridge, Massachusetts, USA*
Wei Wei *Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai, China*
dabo.guan@uea.ac.uk

CORRECTION

An editing error in the Correspondence by A.C. Lewis *et al. (Nature* **526**, 195; 2015) attributed 29,000 UK deaths to particulate matter from diesel, rather than from all sources.

El Niño and intense tropical cyclones

ARISING FROM F.-F. Jin, J. Boucharel & I.-I. Lin *Nature* **516**, 82–85 (2014); doi:10.1038/nature13958

The El Niño/Southern Oscillation (ENSO) influences global climate as well as extreme weather events such as floods, droughts, and tropical cyclones, leading to large societal impacts globally^{1–3}. Jin *et al.*⁴ have shown that El Niño—the warm phase of ENSO—effectively discharges oceanic heat into the central to eastern North Pacific basin through the subsurface ocean after its wintertime peak, resulting in high tropical cyclone activity during the following tropical cyclone peak season in the eastern North Pacific, which has significant implications for seasonal prediction of tropical cyclone activity in the eastern North Pacific. However, we question the robustness of their hypothesis on the following grounds: (1) the correlation between subsurface ocean heat delivered by El Niño and tropical cyclone activity is statistically exaggerated; and (2) wintertime ENSO conditions, which are claimed to have predictive value, are not strongly correlated with tropical cyclone activity during the subsequent summer. These factors imply that the hypothesis of ref. 4 has limitations in practical seasonal tropical cyclone prediction. There is a Reply to this Brief Communication Arising by Jin, F.-F., Boucharel, J. & Lin, I.-I. *Nature* **526**, <http://dx.doi.org/10.1038/nature15547> (2015).

Our first concern is the data smoothing used in their correlation analysis (figure 3 and supplementary figures 10 and 12 of ref. 4). In general, smoothing is a useful tool for removing high-frequency noises in order to focus on long-term variations, but smoothing can also tend to increase correlations artificially (<http://climateaudit.org/2008/02/12/data-smoothing-and-spurious-correlation/>). Jin *et al.*⁴ used a three-year smoothing, which is a suitable technique when the physical variations being examined are multiannual. However, the use of three-year smoothing is not appropriate in this case because ref. 4 examined interannual variations of tropical cyclone activity, focusing on interseasonal connections between wintertime ENSO and summertime tropical cyclones. It turns out that the smoothing significantly increased the correlation between subsurface ocean heat delivered by El Niño (based on the principal component of the second empirical orthogonal function mode in ref. 4, PC2) and tropical cyclone activity (accumulated cyclone energy, ACE⁵) from 0.29 to 0.62 (see Fig. 1a and figure 3b of ref. 4). The smoothing also enhanced the correlation of a bilinear regression model of ref. 4 (see figure 3e of ref. 4) from 0.37 to 0.64 (not shown).

A second concern is their way of comparing tropical cyclone activity between high and low subsurface heat content (PC2 active) periods. Jin *et al.*⁴ showed a surprisingly distinct difference in the number of intense tropical cyclones between the two groups (see figure 2, tables 1 and 2, and supplementary figures 5–9 of ref. 4), but the results were overstated by unequal sampling. Jin *et al.*⁴ sampled 43 months

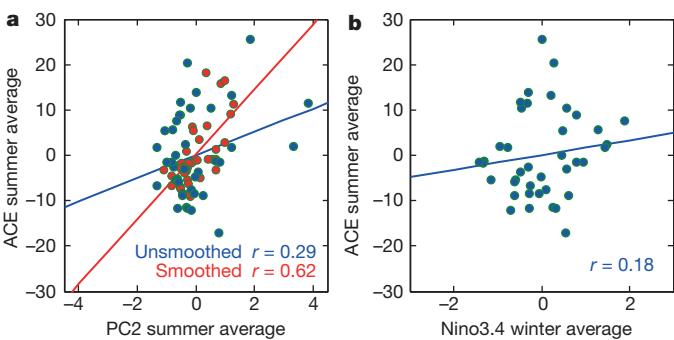


Figure 1 | Relationships of subsurface ocean heat delivered by El Niño (PC2) and the ENSO signal (Niño3.4) with tropical cyclone activity in the eastern North Pacific. **a**, Scatter plot of ACE against PC2 with smoothed (red) and unsmoothed (blue) data. **b**, Scatter plot of ACE against the Niño3.4 index. ACE and PC2 are averaged over the boreal summer and autumn (June to November) and Niño3.4 is averaged over the ENSO peak period (December to February) before the summer tropical cyclone season. Linear regressions (straight lines) and correlation coefficients *r* are shown. All data sets are detrended, and are unsmoothed except for the smoothing discussed above. For consistent comparisons with ref. 4, we used the same analysis period (1970–2009), methods, and data as in ref. 4, except for the smoothing. The subsurface ocean heat over the upper 105 m (T105) represents all the existing subsurface ocean conditions in the eastern North Pacific, whereas PC2 explains only 56% of the T105 variations through a process delivered by El Niño. Therefore, PC2 is the index explaining the relationship between subsurface ocean heat delivered by El Niño and tropical cyclone activity in ref. 4, although T105 is more closely related to tropical cyclone activity than PC2.

for high-heat-content periods, but only sampled 25 months for low-heat-content periods. The monthly distributions of the sampled month were also different. For example, in the climatological lowest (June) and highest (August) tropical cyclone genesis months for the eastern North Pacific season, the numbers of selected months were 8 (19% of total) and 7 (16%) for high-heat-content periods, but 8 (32%) and 2 (8%) for low-heat-content periods, respectively (Table 1). Using this unequal sampling, the total numbers of tropical cyclones are compared for the selected months of the two groups. This is an unfair comparison, leading naturally to a higher number of tropical cyclones in the high-heat-content periods. An impartial comparison should examine the differences in terms of mean values for each month rather than the total number of tropical cyclones. Our monthly comparison with mean values revealed that there are no significant differences in intense tropical cyclone numbers between the high-

Table 1 | Monthly and summer mean numbers of intense tropical cyclones during periods of low and high subsurface heat content

Periods	Total number of selected months	Monthly mean numbers of intense tropical cyclones (number of selected months)						Summer-mean intense tropical cyclone numbers
		June	July	August	September	October	November	
High PC2	43	0.50 (8)	1.38 (8)	1.14 (7)	1.43 (7)	0.86 (7)	0.00 (6)	5.30
Low PC2	25	0.25 (8)	1.14 (7)	1.00 (2)	1.00 (2)	1.00 (2)	0.00 (4)	4.39
Climatology		0.31	0.86	0.90	0.75	0.43	0.00	3.25
Confidence level for difference between periods (%)		67	27	8	31	16	0	67

Intense (category 3, 4 and 5) tropical cyclones represent tropical cyclones at category 3 or above according to the Saffir–Simpson hurricane scale. Input data and methods are the same as in supplementary table 1 of ref. 4 except that we used monthly mean values. Owing to the different monthly distributions in sampling between two periods, summer-mean intense tropical cyclone numbers are calculated by using the sum of monthly mean tropical cyclone numbers, not the total summer tropical cyclone number divided by the total number of selected months. This method has the advantage of not being influenced by the months without tropical cyclones (such as November). High (low) PC2 indicates a period of high (low) subsurface heat content, which is based on the principal component of the second empirical orthogonal function mode in ref. 4.

heat-content and low-heat-content periods based on PC2 (all confidence levels are less than 67%, Table 1).

A third concern is the limitation in the seasonal prediction of tropical cyclone activity. Jin *et al.*⁴ argued that observed ENSO signals (the Niño index) in the winter are good indicators of tropical cyclone activity during the subsequent summer in the eastern North Pacific. However, the correlation between the Niño index in the winter and ACE during the subsequent summer is very low ($r = 0.18$, Fig. 1b), which implies that the subsurface ocean heat delivered by El Niño has very little contribution ($\sim 3\%$) to the total variations of tropical cyclone activity in the subsequent summer. Consequently, ref. 4's theory has limitations in practical seasonal tropical cyclone prediction.

In spite of these concerns, we agree that the delivery of subsurface ocean heat for specific big El Niño events does have an influence on tropical cyclone activity in the eastern North Pacific. In these cases, there is no doubt that ocean thermal control can be used as a predictor for seasonal prediction of tropical cyclones in the eastern North Pacific. However, we consider that ref. 4's claims of a direct connection between the subsurface ocean thermal control delivered by El Niño and tropical cyclone activity are premature for general ENSO events. The connections are not robust enough to apply for the seasonal prediction for all types of ENSO events. An analysis classified according to the types of ENSO (El Niño or La Niña; cold tongue or warm pool El Niño^{6,7}) will be required to improve the correlations and reliability of prediction.

Jin *et al.* reply

REPLYING TO I.-L. Moon, S.-H. Kim & C. Wang *Nature* **526**, <http://dx.doi.org/10.1038/nature15546> (2015)

Observational and modelling studies suggest that subsurface ocean temperature plays a major part in tropical cyclone intensification^{1,2}. In a recent Letter³ we reported that through the El Niño/Southern Oscillation (ENSO) recharge–discharge mechanism, the subsurface heat of ENSO can directly affect intense tropical cyclones in the eastern Pacific^{3,4}. In the accompanying Comment⁵, Moon *et al.* questioned the robustness and relevance of our results.

First, ref. 5 questioned our use of a three-year running mean. However, this simple smoothing is actually physically relevant. It effectively filters out the ‘noise’ from warm pool El Niño events⁶, which do not effectively discharge heat into the eastern Pacific and thus have little control on eastern Pacific tropical cyclone intensity⁷. By focusing on the slow and strong events, we delineated a clear physical mechanism: canonical ENSO events, which mostly have significant discharge of heat, exert a strong oceanic control on tropical cyclone activity in the eastern Pacific. However, we acknowledge that we should have explicitly indicated this point in ref. 3 to avoid confusion. Furthermore, it should be emphasized that although not every ENSO event delivers a great amount of heat to the eastern Pacific region, subsurface heat anomalies do have a strong impact on the overlying tropical cyclone activity. Using unfiltered data (that is, without the three-year running mean), a strong correlation ($r = 0.62$) between monthly subsurface ocean heat over the top 105 m (T105) of depth and monthly tropical cyclone intensity is still evident (Fig. 1), clearly supporting a strong relationship between subsurface heat and the eastern Pacific tropical cyclone intensity. We also note that the quality of both tropical cyclone and ocean data sets is problematic in the earlier period (1959–1978), as there exist few subsurface observations⁸, and there were no systematic satellite tropical cyclone observations before the 1980s⁹. Thus, these low-quality data sets can

Il-Ju Moon¹, Sung-Hun Kim¹ & Chunzai Wang²

¹Typhoon Research Center, Jeju National University, 102 Jejudaehak-ro, Jeju 690-756, South Korea.

email: ijmoon@jejunu.ac.kr

²NOAA/Atlantic Oceanographic and Meteorological Laboratory, Miami, Florida 33149, USA.

Received 22 March; accepted 28 July 2015.

1. Wang, C. & Lee, S.-K. Is hurricane activity in one basin tied to another? *Trans. AGU* **91**, 93–94 (2010).
2. Chan, J. C. L. Tropical cyclone activity over the western North Pacific associated with El Niño and La Niña events. *J. Clim.* **13**, 2960–2972 (2000).
3. Pielke, R. A. Jr & Landsea, C. W. La Niña, El Niño, and Atlantic Hurricane Damages in the United States. *Bull. Am. Meteorol. Soc.* **80**, 2027–2033 (1999).
4. Jin, F.-F., Boucharel, J. & Lin, I.-I. Eastern Pacific tropical cyclones intensified by El Niño delivery of subsurface ocean heat. *Nature* **516**, 82–85 (2014).
5. Maru, R. N. Recent historically low global tropical cyclone activity. *Geophys. Res. Lett.* **38**, L14803 (2011).
6. Yeh, S.-W. *et al.* El Niño in a changing climate. *Nature* **461**, 511–514 (2009).
7. Kug, J.-S., Jin, F.-F. & An, S.-I. Two types of El Niño events: cold tongue El Niño and warm pool El Niño. *J. Clim.* **22**, 1499–1515 (2009).

Author Contributions I.-J.M. conceived the idea, designed the study, and wrote the Comment. S.-H.K. conducted most of the analysis and discovered main results. C.W. contributed to the interpretation of the results and editing of the manuscript.

Competing Financial Interests Declared none.

doi:10.1038/nature15546

contribute to lower correlations in the earlier period (Fig. 1b). In summary, we do not agree with the contention of ref. 5 that the impacts of subsurface thermal control on tropical cyclone activity are largely exaggerated owing to the smoothing regardless of the period.

Second, ref. 5 questioned the statistical significance of tropical cyclone activity changes between periods of high and low subsurface heat over the eastern Pacific region. Our original choice (ref. 3) of a broad temporal measure in terms of simple accounts of tropical cyclone number per decade displays significant differences between these periods (periods of high and low subsurface heat are referred to as high PC2/low PC2; see ref. 3 for details). Moon *et al.*⁵ argue that this significance is severely degraded when their ‘accurate’ counting of total number per month is used. However, we believe that their counting ignored one important fact: there are many months without any tropical cyclone (hurricane) occurrence in the record. We argue that those ‘hurricane-empty’ months should be removed for a truly accurate counting. If we use the

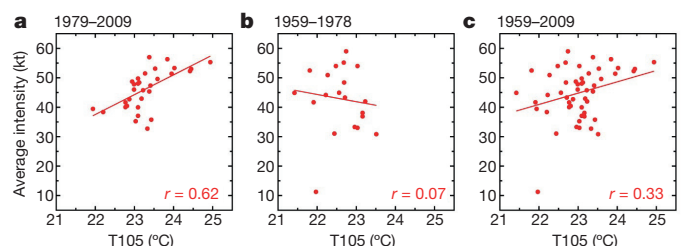


Figure 1 | Relationships between heat content and tropical cyclone intensity in the northeast Pacific. a, Correlation based on the unsmoothed data from 1979–2009. b, Correlation based on the earlier period (1959–1978). c, Correlation based on the entire period (1959–2009).

Table 1 | Changes in tropical cyclones statistics between high- and low-heat-content periods in the northeast Pacific

	Number of sampled month	Mean of the number of category 3–5 tropical cyclones	Confidence level (two-tailed)	Confidence level (one-tailed)
a	High PC2	29	1.34	93%
	Low PC2	21	0.76	96.5%
	Mean (%)	Variance	Number of months in sample	t value (confidence level)
b	High PC2	11.24	34	3.368 (>99.75% for one-tailed; >99.5% for two-tailed)
	Low PC2	3.21	24	

a, Comparison of intense (category 3–5) tropical cyclone count between high and low PC2 active periods, using category 1 as a minimum threshold.

b, t-test result for the percentage of intense tropical cyclone grid count—that is, grid count of category 3–5 tropical cyclones divided by the grid count of total tropical cyclones (from tropical depression to category 5)—in the tropical cyclone region.

conventional definition of tropical cyclone (category 1 or above) to filter out tropical depressions and storms that have little interactions with the subsurface ocean, and then remove all ‘hurricane-empty’ months from the counting, the number of tropical cyclones per month again passes the 95% confidence level (Table 1a). Furthermore, using a stricter grid-by-grid approach, a precise measure of intense tropical cyclone activity giving the percentage of intense tropical cyclones over all tropical cyclone grids, we obtained an even stronger statistical significance (above the 99.5% confidence levels, see Table 1b). Both the broad measure in our earlier analysis and the stricter measures we present here give consistent results regarding the significance of this statistical test.

Third, Moon *et al.*⁵ questioned the usefulness of our finding in the prediction of tropical cyclones. It is known (and expected to continue to be so) that a major El Niño event is followed by a highly predictable heat discharge into the eastern Pacific. For instance, as this year’s El Niño develops into a strong event likely to peak in the coming boreal winter in the equatorial eastern Pacific, there is a high chance (75%) of a significant heat discharge into the eastern north Pacific to fuel intense tropical cyclones when next year’s tropical cyclone season arrives. Should a weak warm pool El Niño event occur, it would discharge little heat to affect the eastern Pacific tropical cyclone activity. Information about each El Niño and its heat content discharge is known long before the Northern Hemisphere tropical cyclone season, so it has a clear predictive value.

In conclusion, we reaffirm the physical mechanism evidenced by ref. 3, that is, the impact of ocean heat content discharged from major El Niño events on the eastern Pacific tropical cyclone intensity and its value for tropical cyclone predictions beyond the seasonal timescale.

Methods

This work uses single-tailed significance while ref. 5 uses double-tailed. If the mean of the first group is expected to be larger than the second, a single-tailed

approach is used. However, if the mean of the first group is different from the second in either direction, a double-tailed approach is used. In the context of this study, we expect the group 1 (high PC2) mean to be larger than the group 2 (low PC2) mean, so we used a single-tail approach. In the revised Table 1, we present results using both approaches.

F.-F. Jin¹, J. Boucharel² & I.-I. Lin³

¹University of Hawaii at Manoa, Honolulu, Hawaii 96822, USA.

email: jff@hawaii.edu

²ARC Centre of Excellence for Climate System Science, University of New South Wales, Sydney, New South Wales 2052, Australia.

email: bouch@hawaii.edu

³Department of Atmospheric Sciences, National Taiwan University, Taipei 10617, Taiwan.

1. Lin, I. I. *et al.* An ocean coupling potential intensity index for tropical cyclones. *Geophys. Res. Lett.* **40**, 1878–1882 (2013).
2. Vincent, E. M., Emanuel, K. A., Lengaigne, M., Vialard, J. & Madec, G. Influence of upper-ocean stratification interannual variability on tropical cyclones. *J. Adv. Model. Earth Syst.* **6**, 680–699 (2014).
3. Jin, F.-F., Boucharel, J. & Lin, I. I. Eastern Pacific tropical cyclones intensified by El Niño delivery of subsurface ocean heat. *Nature* **516**, 82–85 (2014).
4. Jin, F. F. An equatorial ocean recharge paradigm for ENSO. Part I: conceptual model. *J. Atmos. Sci.* **54**, 811–829 (1997).
5. Moon, I.-J., Kim, S.-H. & Wang, C. El Niño and intense tropical cyclones. *Nature* **526**, <http://dx.doi.org/10.1038/nature15546> (2015).
6. Kug, J.-S., Jin, F.-F. & An, S.-I. Two types of El Niño events: cold tongue El Niño and warm pool El Niño. *J. Clim.* **22**, 1499–1515 (2009).
7. Ren, H. L. & Jin, F.-F. Recharge oscillator mechanisms in two types of ENSO. *J. Clim.* **26**, 6506–6523 (2013).
8. Balmaseda, M. A., Trenberth, K. E. & Källén, E. Distinctive climate signals in reanalysis of global ocean heat content. *Geophys. Res. Lett.* **40**, 1754–1759 (2013).
9. Landsea, C. W. Meteorology, hurricanes and global warming *Nature* **438**, E11–E12 (2005).

doi:10.1038/nature15547

IMMUNOLOGY

Caspase target drives pyroptosis

Inflammatory caspase proteins help to control pathogen replication by triggering pyroptotic cell death. It now emerges that cleavage of the caspase substrate gasdermin D is sufficient to induce pyroptosis. [SEE ARTICLES P.660 & P.666](#)

PETR BROZ

Inflammation and pyroptosis, a specialized form of cell death, are key responses of the innate immune system to pathogens. To avoid excessive damage to the host organism, pyroptosis is tightly regulated by the activation of inflammatory caspase proteins, such as caspase-1 and caspase-11 — but how these caspases initiate this cell-death program has been unknown. In this issue, Shi *et al.*¹ (page 660) and Kayagaki *et al.*² (page 666) show that caspase-mediated cleavage of the protein gasdermin D creates an amino-terminal fragment that has intrinsic pyroptosis-inducing activity.

Caspases are a family of enzymes best known for controlling apoptosis, a well-characterized type of cell death that is vital for embryonic development and tissue homeostasis in adults. A subgroup, comprising caspase-1, -11 and -12 in mice and caspase-1, -4, -5 and -12 in humans, controls the inflammatory response to pathogens and noxious stimuli in the cytoplasm of host cells. These inflammatory caspases are produced as inactive monomers and become activated when recruited to multi-protein complexes known as inflammasomes³. ‘Canonical’ inflammasomes activate caspase-1 and are assembled by cytoplasmic sensors, such as the NOD-like receptors, which detect diverse pathogen- and host-derived danger signals. By contrast, the bacterial cell-wall component lipopolysaccharide (LPS), one of the strongest immune-system activators, leads to the assembly of ‘non-canonical’ inflammasomes⁴ through direct binding and activation of caspase-4, -5 or -11.

Although the signals leading to inflammasome assembly are quite well understood, the downstream signalling events that follow the activation of inflammatory caspases are poorly characterized. Initial work identified the pro-inflammatory cytokine protein IL-1 β as a key substrate of caspase-1 in the canonical inflammasome pathway. Subsequent studies discovered that inflammatory caspases also induce a programmed cell-death pathway, which involves cell swelling, lysis and the release of cytoplasmic content, presumably as a result of the formation of membrane pores, and is thus morphologically distinct from apoptosis.

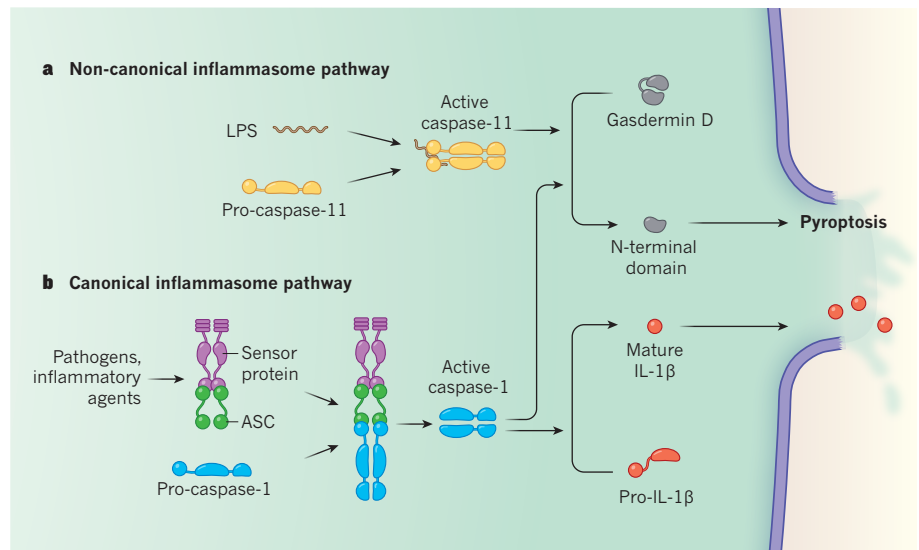


Figure 1 | Gasdermin D cleavage initiates pyroptosis during inflammasome activation. **a**, The non-canonical inflammasome pathway detects bacterial lipopolysaccharide (LPS) molecules in the cytoplasm of infected cells. Direct binding of LPS to the protein pro-caspase-11 causes the protein's dimerization, to become active caspase-11. In human cells, caspase-4 and -5 perform the function of mouse caspase-11. Shi *et al.*¹ and Kayagaki *et al.*² show that caspase-11 cleaves the protein gasdermin D, releasing its amino-terminal domain from inhibition by its carboxy-terminal domain. The N-terminal domain (directly or indirectly) drives pyroptosis, a cell-death process that causes cell lysis and the release of cytoplasmic content. **b**, The canonical inflammasome pathway is triggered by different cytoplasmic sensor proteins, which detect pathogens and inflammatory agents. These recruit pro-caspase-1 monomers through the adaptor protein ASC and activate the caspase by dimerization. The findings of the two groups suggest that caspase-1 can also initiate pyroptosis by cleaving gasdermin D, although other pyroptosis-inducing caspase-1 substrates may exist. Caspase-1 also processes the pro-inflammatory cytokine pro-IL-1 β to generate mature IL-1 β , which is presumably released by cell lysis during pyroptosis.

Because this type of cell death is intrinsically pro-inflammatory, it was named pyroptosis, from the Greek *pyro* (fire or fever) and *ptosis* (to fall)⁵. Yet despite the identification of many potential caspase-1 substrates, the key molecules involved in pyroptosis have remained a mystery.

Now, Shi *et al.* and Kayagaki *et al.* have uncovered how inflammatory caspases induce pyroptosis. Taking advantage of the development of CRISPR/Cas9 genome-editing methods, Shi *et al.* used a genome-wide screen to identify host factors involved in caspase-11- and caspase-1-induced pyroptosis. Kayagaki *et al.* took a different tack, using mice subjected to random mutation (through ENU mutagenesis) to find mediators of caspase-11-dependent non-canonical inflammasome signalling. Strikingly, both groups

identified the same gene, *Gsdmd*, which encodes gasdermin D, a member of the poorly characterized gasdermin protein family. They confirmed their screening results through several experimental approaches, demonstrating that gasdermin D is essential for pyroptosis induced by non-canonical inflammasome activation in mouse and human cells (Fig. 1a). Kayagaki *et al.* also show that *Gsdmd* deficiency reduces LPS-induced lethality in mice, which is known to depend on caspase-11-induced pyroptosis.

Although both groups find that gasdermin D is essential for pyroptosis induced by caspase-4, -5 or -11, their results diverge on the requirement of gasdermin D for pyroptosis resulting from caspase-1 activation (Fig. 1b). Confirming the results of their CRISPR/Cas9 screen that identified *Gsdmd* as an essential

gene for caspase-1-induced pyroptosis, Shi and colleagues found that pyroptosis is abolished in *Gsdmd*-deficient macrophages (a type of immune cell) exposed to activators of canonical inflammasomes for short times. Kayagaki *et al.* present similar results; but after prolonged incubation, which may better reflect physiological processes, they no longer detected differences between the pyroptotic responses of wild-type and *Gsdmd*-deficient macrophages. For now, the jury is still out on whether gasdermin D is an essential mediator of caspase-1-induced pyroptosis; if it is not, then other pyroptosis-inducing caspase-1 substrates must await discovery.

Another enigma in the inflammasome field is the release of mature IL-1 β after it has been cleaved by caspase-1. IL-1 β lacks the amino-acid sequence that would direct it to the Golgi apparatus⁶, a cellular substructure that mediates protein secretion, so it is thought to be released by an unconventional secretion mechanism. Shi *et al.* show that, in *Gsdmd*-deficient macrophages, caspase-1 activity and IL-1 β processing were not affected after canonical inflammasome activation, but secretion of the mature cytokine was completely abrogated. These results strongly support the hypothesis that IL-1 β is released following pyroptotic lysis of the cell.

How do inflammatory caspases activate gasdermin D to induce pyroptosis? Consistent with the previous identification of the protein as a caspase-1 substrate⁷, both groups noted that gasdermin D is cleaved by caspase-1 or caspase-11 at the aspartate 276 amino-acid residue following canonical or non-canonical inflammasome activation, and that mutation of this cleavage site completely abolished pyroptosis. The N-terminal domain that was left after cleavage was sufficient to induce cell death showing morphological features of pyroptosis, and the carboxy-terminal domain was found to bind the N-terminal domain if the former was overexpressed, blocking cell death. These results are consistent with the hypothesis that caspase-dependent cleavage releases the N-terminal domain of gasdermin D from an inhibitory interaction with its C-terminal domain, thereby unleashing its cell-killing properties.

Finally, the studies highlight the role of other gasdermin family members in pyroptosis. In mice, gain-of-function mutations in *Gsdma3* cause skin thickening and hair loss that is associated with chronic skin inflammation⁸. Shi *et al.* show that these mutations abrogate the interaction between the C- and N-terminal domains of gasdermin A3, and that the N-terminal domain of this protein also induces pyroptosis. Because only gasdermin D, and no other gasdermins, features a caspase cleavage site, further research is necessary to understand the context and mechanism of this seemingly similar activation process.

The identification of gasdermin D as a

key mediator of pyroptosis is a conceptual advance in our understanding of how inflammatory caspases cause cell death. The studies present compelling evidence that caspase-induced cleavage of a single substrate, and not a spectrum of different substrates, is sufficient to induce pyroptosis and all the associated morphological features — similar to the action reported for two substrates of apoptotic caspases, ATP11C (ref. 9) and ROCK1 (refs 10, 11). Precisely how the gasdermin D N-terminal domain exerts its cytotoxic function should be addressed in future studies. Does it simply relay the signal to yet-unknown pyroptosis inducers, or might the fragment itself be an executioner of cell death, like the pseudokinase protein MLKL, for example, which drives necroptotic cell death by forming pores in the cell membrane¹²? Elucidating the mechanism of gasdermin D's cytotoxic activity is likely to bring further exciting discoveries, and could pave the way for therapeutic approaches for

inflammasome-associated inflammatory and metabolic disorders. ■

Petr Broz is at the Biozentrum, University of Basel, CH-4056 Basel, Switzerland.
e-mail: petr.broz@unibas.ch

1. Shi, J. *et al.* *Nature* **526**, 660–665 (2015).
2. Kayagaki, N. *et al.* *Nature* **526**, 666–671 (2015).
3. Martinon, F., Burns, K. & Tschopp, J. *Mol. Cell* **10**, 417–426 (2002).
4. Shi, J. *et al.* *Nature* **514**, 187–192 (2014).
5. Cookson, B. T. & Brennan, M. A. *Trends Microbiol.* **9**, 113–114 (2001).
6. Rubartelli, A., Cozzolino, F., Talio, M. & Sitia, R. *EMBO J.* **9**, 1503–1510 (1990).
7. Agard, N. J., Maltby, D. & Wells, J. A. *Mol. Cell. Proteomics* **9**, 880–893 (2010).
8. Ruge, F. *et al.* *J. Invest. Dermatol.* **131**, 572–579 (2011).
9. Segawa, K. *et al.* *Science* **344**, 1164–1168 (2014).
10. Coleman, M. L. *et al.* *Nature Cell Biol.* **3**, 339–345 (2001).
11. Sebbagh, M. *et al.* *Nature Cell Biol.* **3**, 346–352 (2001).
12. Cai, Z. *et al.* *Nature Cell Biol.* **16**, 55–65 (2014).

This article was published online on 16 September 2015.

PHOTONICS

Random sudoku light

A clever approach has been used to imprint a phase pattern on a laser beam. The pattern is not only random at each point, but also depends on information stored elsewhere in the pattern.

TONI EICHELKRAUT & ALEXANDER SZAMEIT

Imagine that you are stuck in traffic. What determines your options? Only the car directly in front of you: if it travels slowly, you drive slowly; if it stops, so must you. However, your decisions are independent of all other drivers' decisions. This observation has profound physical consequences, because it leads to the distinction between two classes of random physical system, known as Markovian and non-Markovian after the mathematician Andrey Markov¹. Traffic jams are Markovian systems, and, so far,

this has also been the case for light that has a random spatial distribution in intensity or phase. Writing in *Physical Review Letters*, Fischer *et al.*² report the first experimental demonstration of non-Markovian light.

The behaviour of each member of a Markovian system is determined only by its nearest neighbours. By contrast, each member of a non-Markovian system depends on the behaviour of a larger region, or even of the

entire system. Light fields that have random spatial distributions of phase have so far been generated exclusively as the Markovian type, using processes such as scattering in diffusive optical media³. Fischer and colleagues use a clever phase-imprinting approach to form a light field that has a fully random phase at each spatial point and a local random nature that is determined not only by the nearest-neighbouring points, but also by regions that are farther away.

Markov originally formulated his groundbreaking theory with dynamically evolving, random systems in mind. He formalized the treatment of an important feature of such systems: how a system's 'memory' of its previous states affects its future states. Markov realized that, in a system that has no memory, future events can be predicted solely on the basis of the system's present state: the defining property of a Markovian process is that the past and the future are mutually independent. Examples include the motion of dust floating in the air, a sequence of simple coin tosses or the PageRank algorithm that Google uses to search the Internet. The defining property of a Markovian process can be generalized for systems that have random spatial distributions, the traffic jam being one of the simplest examples.

As an example of a random non-Markovian

5	9	4	3	2	8	7	6	1	2	3	9	5	8	4	9
1	3	8	6	7	5	9	4	2	8	5	7	6	1	3	8
6	7	2	4	1	9	5	8	3	6	4	1	7	2	9	3
7	6	5	2	3	4	8	1	9	4	6	5	3	7	2	1
8	4	9	7	5	1	2	3	6	9	7	8	1	4	5	7
3	2	1	8	9	6	4	5	7	3	1	2	9	6	8	4
9	8	6	1	4	7	3	2	5	7	8	6	4	9	1	6
4	5	3	9	6	2	1	7	8	5	9	4	2	3	6	5
2	1	7	5	8	3	6	9	4	1	2	3	8	5	7	2
3	6	1	7	9	5	4	8	2	6	3	1	9	7	5	1
8	9	2	3	1	4	7	5	6	8	4	9	3	1	2	4
7	4	5	8	2	6	9	3	1	2	5	7	6	8	4	7
1	3	4	2	7	8	5	6	9	4	7	8	1	2	3	8
6	7	8	4	5	9	2	1	3	9	6	5	7	4	8	9
5	2	9	6	3	1	8	4	7	3	1	2	5	6	9	3
7	6	5	2	3	4	8	1	9	4	6	5	3	7	2	1

Figure 1 | Non-Markovian light. Fischer *et al.*² used overlapping sudoku puzzles to imprint a spatially non-Markovian phase pattern on laser light using a device known as a spatial light modulator (SLM). The authors filled the cells in the upper left square of the SLM with phase information corresponding to the numbers of a correctly solved 9×9 sudoku (blue). They then shifted the puzzle frame down (or to the right) by six lattice cells, and filled it with a phase pattern corresponding to a sudoku solution for the new position, which incorporates the numbers from the bottom three rows (or rightmost three columns) of the first solution; new solutions are shown in pink. In this process, the number found in each sudoku cell depends not only on those in the nearest cells, but also on those in all other cells of the corresponding sudoku — that is, the pattern is non-Markovian. This is in contrast to the Markovian pattern of a fully random distribution of numbers (not shown).

process, think of an urn that contains two red balls and one blue ball. On three consecutive days, one ball is drawn and not replaced. If yesterday a red ball was drawn and today a blue ball is drawn, then it is clear that tomorrow a red ball will be drawn. We know this for certain because we have taken into account the events that have affected the urn's state today and yesterday; that is, tomorrow's draw depends not only on today's draw, but also on past draws.

To generate light that has a non-Markovian spatial distribution, Fischer *et al.* modified the transverse phase pattern of a laser beam using a device known as a spatial light modulator (SLM; an array of liquid-crystal cells akin to a computer screen). The authors then imprinted on the laser beam a particular phase pattern that corresponds to numerical solutions to overlapping sudoku puzzles; these were chosen because they are non-Markovian systems (Fig. 1).

Fischer and colleagues started this process by filling the upper left square of the SLM array with phase information corresponding to the numbers of a correctly solved 9×9 sudoku. They then shifted the boundaries of the sudoku frame by a certain number of cells, either vertically or horizontally, and solved a new puzzle at the new grid position, adjusting the SLM's phase accordingly. This procedure was repeated until the entire SLM array was filled with phase information that had a random, but non-Markovian, distribution.

The phase imprint of light generated in this way is non-Markovian because the information contained in a given SLM cell depends not only on the contents of the nearest cells, but also on those of all other cells of the corresponding sudoku

PLANT BIOLOGY

Pigments on the move

In plant cells, the pigment anthocyanin is transported to a membrane-bounded organelle called the vacuole for storage. A previously unidentified transport pathway involving vacuolar-membrane extensions mediates this process.

DIANE C. BASSHAM

Pigment molecules called anthocyanins give plant cells a dark red or purple colour that attracts pollinating insects to flowers¹ and protects leaves against damage from ultraviolet light². Anthocyanins are stored in intracellular structures called anthocyanin vacuolar inclusions (AVIs) within a large organelle called the vacuole, the environment of which protects the pigments from oxidation and subsequent loss of colour. There has been much debate^{3–5} about how anthocyanins are transported to AVIs from their site of synthesis in another organelle, the endoplasmic

reticulum. A remarkable aspect of the authors' implementation of this process is that they used a tuning parameter to control the number of cells by which the sudoku frames can be shifted. Using this parameter, the authors investigated the transition between periodic and non-Markovian light, by tuning the imprinted phase pattern from one that corresponds to strictly periodic (largely overlapping) sudokus to random sequences of independent puzzles.

Fischer and co-workers have shown that light that has a non-Markovian spatial distribution can be generated in a highly flexible manner using a simple and fast set-up. These fascinating results pave the way to generating new types of 'patterned' light beam, in which properties such as intensity or polarization are used to store information. The authors' technique might help scientists to achieve a deeper understanding of non-Markovian processes. Moreover, the 'sudoku approach' could serve as a general strategy for building optical or other systems that have the singularly defining non-Markovian property, spatial memory. ■

Toni Eichelkraut and Alexander Szameit
are at the Institute of Applied Physics,
Abbe Centre of Photonics, University of Jena,
07749 Jena, Germany.
e-mail: alexander.szameit@uni-jena.de

1. Markov, A. A. *Proceedings of the Physical and Mathematical Society at the University of Kazan* 2nd edn, Vol. 15, 135–156 (1906).
2. Fischer, R. *et al.* *Phys. Rev. Lett.* **115**, 073901 (2015).
3. Fairchild, M. D. *Color Appearance Models* 2nd edn, 65 (Wiley, 2005).

reticulum. Writing in *The Plant Cell*, Chanoca *et al.*⁶ describe a previously unknown pathway of intracellular transport in which extensions of the membrane that surrounds the vacuole engulf and internalize anthocyanins.

Several models for anthocyanin transport have previously been proposed. One model³ states that anthocyanins are carried directly across the vacuolar membrane by transport proteins, with the help of glutathione S-transferase enzymes. A second⁴ proposes that anthocyanins first translocate into the interior of the endoplasmic reticulum, a small part of which then pinches off to form membrane-bounded vesicles that transport the pigment

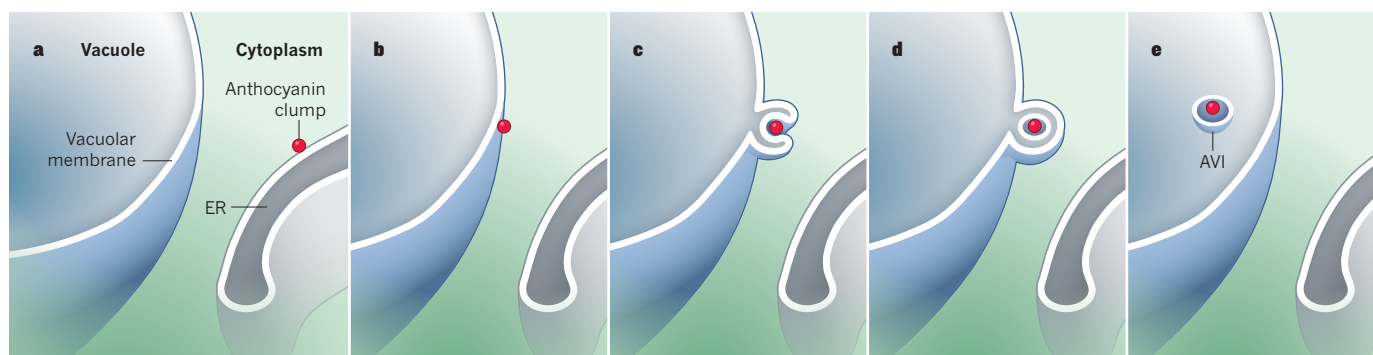


Figure 1 | Transport of plant pigments. **a**, In plant cells, pigment molecules called anthocyanins are synthesized in an organelle called the endoplasmic reticulum (ER). Chanoca *et al.*⁶ propose a mechanism by which anthocyanins are transported to another organelle, the vacuole, for storage. **b**, In their model, clumps of anthocyanin move from the ER across the

cytoplasm and become tightly pressed against the vacuole. **c**, The membrane that surrounds the vacuole extends around the clump. **d**, The membrane then fuses, creating a single-layered membrane structure around the pigment. **e**, This structure, called an anthocyanin vacuolar inclusion (AVI), is then released into the vacuole.

and expel it into the vacuole, analogous with the well-studied mechanism by which proteins are transported to the vacuole.

A third model⁵ involves an autophagy-like pathway. Autophagy is the process by which unwanted materials are transported from the cytoplasm into the vacuole for degradation, and autophagy-related pathways can also transport materials for storage⁷. In one type of autophagy, dubbed macroautophagy, vesicles called autophagosomes, which have a double membrane, form around the cytoplasmic cargo. The outer autophagosome membrane fuses with the vacuolar membrane, depositing the inner membrane and the enclosed cargo into the vacuolar interior. In this model, the engulfment of part of the endoplasmic reticulum and its associated anthocyanins by autophagosomes leads to their deposition inside the vacuole.

Chanoca *et al.* attempted to distinguish between these models using fluorescence microscopy to study cells harbouring a fluorescent dye that labels membranes, and transmission electron microscopy to study the process at high resolution. They assessed the structure of AVIs in leaf cells from the model plant *Arabidopsis thaliana* and in cells from the purple petals of *Eustoma grandiflorum* (commonly known as lisianthus). The first two models of anthocyanin transport predict that AVIs would not be bounded by membranes, whereas the third model predicts that the AVIs would have two membranes — one derived from the autophagosome inner membrane and the other from the engulfed endoplasmic reticulum. Surprisingly, the authors found that the AVIs were surrounded by a single membrane, compelling them to rethink current models.

To explain their findings, Chanoca and colleagues again turned to microscopy to analyse plant cells that produce a fluorescent protein that labels the vacuolar membrane. This strategy allowed the authors to reconstruct the events leading to the vacuolar uptake of anthocyanins. They observed that clumps

of anthocyanin in the cytoplasm outside the endoplasmic reticulum initially become tightly pressed against the vacuolar membrane, which puts out extensions that wrap around the anthocyanins. These extensions finally fuse, releasing anthocyanins surrounded by a single, vacuole-derived membrane into the vacuole (Fig. 1). This mechanism is reminiscent of a process called microautophagy — in which the vacuole engulfs cytoplasmic material directly — rather than of the indirect macroautophagy route.

Although there has been speculation that microautophagy occurs in plants⁸, mechanistic information is lacking. In other organisms, microautophagy, macroautophagy and similar processes typically depend on a group of core autophagy (ATG) genes⁹. To assess the relationship between anthocyanin uptake and known autophagy pathways, the authors analysed the formation of AVIs in cells that do not express the gene *ATG5*, and in cells treated with a drug that blocks autophagy. AVIs formed normally in both cases, indicating that, despite the morphological similarities, anthocyanin uptake and autophagy are distinct at the molecular level. The anthocyanin-uptake pathway therefore seems to be a previously unidentified type of mechanism for vacuolar uptake of material. It should be noted, however, that this finding does not discredit the other models of anthocyanin transport, because there could be more than one uptake mechanism at work.

The molecular pathways that direct deformation of the vacuolar membrane and so promote anthocyanin engulfment remain to be determined. It is possible that a subset of ATG proteins not studied by Chanoca *et al.* is required, and analysis of other mutant plants could address this issue. However, it seems more likely that a different mechanism is responsible for the generation of membrane protrusions from the vacuole, potentially involving proteins that induce membrane curvature.

How anthocyanins are selected for

engulfment over other nearby cytoplasmic components is also an open question. One possibility is that an anthocyanin-binding protein on the vacuolar membrane concentrates the pigment at sites of uptake. Alternatively, given that some evidence¹⁰ suggests that anthocyanins and related molecules bind to lipids, the pigments might bind directly to the lipid bilayer that makes up the membrane, possibly even causing membrane curvature at the binding site.

Other anthocyanin-related molecules — including tannins, which protect plants from being eaten, and possible photoprotectants called coumarin glucosides — also accumulate in vacuolar inclusions, and the mechanisms by which they are taken up by the vacuole are unknown. It could be, then, that Chanoca and colleagues have uncovered a widespread pathway for the transport of pigments and related molecules into the vacuole. Their pathway might yet prove to be crucial beyond its role in pigment accumulation, for plant protection against herbivores and pests, and for abiotic stress tolerance. ■

Diane C. Bassham is in the Department of Genetics, Development and Cell Biology, and at the Plant Sciences Institute, Iowa State University, Ames, Iowa 50011-3650, USA. e-mail: bassham@iastate.edu

1. Iwashina, T. *Nat. Prod. Commun.* **10**, 529–544 (2015).
2. Tohge, T., Kusano, M., Fukushima, A., Saito, K. & Fernie, A. R. *Plant Signal. Behav.* **6**, 1987–1992 (2011).
3. Saito, K. *et al. Plant Physiol. Biochem.* **72**, 21–34 (2013).
4. Poustka, F. *et al. Plant Physiol.* **145**, 1323–1335 (2007).
5. Pourcel, L. *et al. Mol. Plant* **3**, 78–90 (2010).
6. Chanoca, A. *et al. Plant Cell* <http://dx.doi.org/10.1105/tpc.15.00589> (2015).
7. Liu, Y. & Bassham, D. C. *Annu. Rev. Plant Biol.* **63**, 215–237 (2012).
8. van Doorn, W. G. & Papini, A. *Autophagy* **9**, 1922–1936 (2013).
9. Feng, Y., He, D., Yao, Z. & Klionsky, D. J. *Cell Res.* **24**, 24–41 (2014).
10. Sengupta, B., Banerjee, A. & Sengupta, P. K. *FEBS Lett.* **570**, 77–81 (2004).

ECOLOGY

Foraging further

King penguins on the Crozet archipelago in the southern Indian Ocean travel south to forage for food around the Antarctic Polar Front, where cold Antarctic waters meet warmer sub-Antarctic seas (pictured, a king penguin diving). Writing in *Nature Communications*, Bost *et al.* report that climatic variability can alter the birds' foraging behaviour and population dynamics (C. A. Bost *et al.* *Nature Commun.* <http://dx.doi.org/10.1038/ncomms9220>; 2015).

By tracking king penguins (*Aptenodytes patagonicus*) for 16 years, Bost *et al.* found that changes associated with an increased sea surface temperature of just 1 °C pushed the polar front southward, and increased both the distances penguins travelled to forage and the depths to which they dived for food. After large-scale climatic anomalies, their population size also fell. Climate models predict that the front will continue to shift southward, which may threaten penguins and their prey. **Jennifer R. Gardiner**



ANTOINE JORIS

CHEMICAL BIOLOGY

Protein modification in a trice

Organometallic reagents have been developed that chemically modify proteins and peptides specifically at cysteine amino-acid residues — potentially offering a general route to making therapeutically useful compounds. [SEE LETTER P.687](#)

HEATHER MAYNARD

Proteins are valuable therapeutic and imaging agents, but they often need to be chemically modified to work optimally in these roles. If a protein is to remain fully biologically active after a chemical group has been attached, the modification must often take place at a specific, predetermined amino-acid side chain. On page 687 of this issue, Vinogradova *et al.*¹ report palladium-containing organometallic reagents that not only modify proteins and peptides at selective sites, but also do so quickly and without being heated, to generate high yields of products.

Chemically modified proteins can have several advantages over their naturally occurring counterparts. For example, attaching polymers to a protein can increase the time that the protein spends in the body, leading to

fewer injections for patients^{2,3}. Proteins that target certain cancerous tissues can be visualized *in vivo* if radioactive labels are attached, facilitating identification of prognosis and treatment monitoring⁴. And anticancer drugs attached to protein antibodies can be as effective as the unmodified drugs, but have fewer side effects⁵.

Therapeutic proteins have conventionally been modified non-selectively, but this sometimes leads to large reductions in biological activity — as much as 93% when polymers are attached⁶. However, site-specific modifications can be crucial for the modified product to retain full biological activity. Many organic reagents have been used to modify proteins at specific sites, with varying success. By contrast, the use of organometallic reagents has lagged behind that of their wholly organic counterparts because of the difficulty of finding

reaction conditions that are compatible with the use of both proteins and organometallic reagents. Furthermore, it is hard to achieve selectivity in the presence of the many reactive chemical groups found in proteins⁷.

One approach to achieving site selectivity is to exploit either naturally occurring or non-natural amino acids that have unique reactivity, by judiciously placing them at protein sites at which the attachment of a chemical group will not alter the protein's function or structure. Cysteine is commonly used for this purpose, because the thiol group (SH) in this natural amino acid's side chain can be targeted selectively depending on the reagent or the pH used in the modification reaction. Many proteins do not have 'free' cysteines that contain thiols, because these groups often react with other cysteine residues in the same protein to form disulfide bonds (S–S). But modern biochemical techniques now allow a free cysteine to be placed in any position in a protein's amino-acid sequence, thereby pre-directing the site of attachment.

Organometallic approaches to modifying proteins^{7,8} have most often targeted non-natural amino-acid residues or certain natural ones (lysine, tyrosine or tryptophan residues), but rarely free cysteines. In Vinogradova and colleagues' method, the organometallic reagent reacts specifically with the thiol groups of free cysteines across a broad pH range (5.5–8.5). The reactions are complete within minutes, and can be performed using low

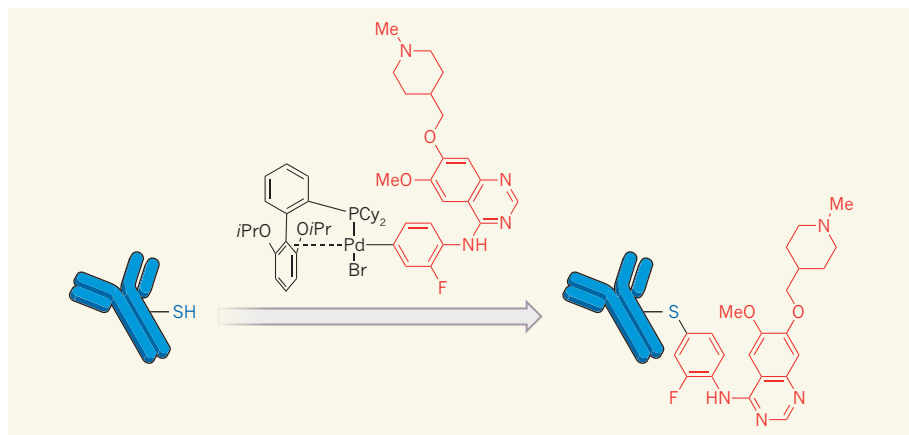


Figure 1 | Site-selective reactions with proteins. Vinogradova *et al.*¹ describe organometallic reagents that react selectively with thiol groups (SH) in the side chains of cysteine amino-acid residues in proteins. Here, a thiol group in an antibody protein (blue) reacts with a palladium complex that contains the anticancer drug vandetanib (red), to produce an antibody–vandetanib conjugate; on average, 4.4 thiols in the antibody were modified. The reactions proceed in high yields, at room temperature, to form stable bonds within minutes.

concentrations (micromolar concentrations or less) of a protein or peptide — a useful attribute because these biomolecules are often available in only small quantities. Remarkably, the reaction yields are normally 100%, and the resulting bonds are chemically robust.

The authors report that their reactions work in aqueous solutions (which are required for proteins), with the addition of a small amount of organic solvent. The transformations also work in air, allowing them to be set up easily on the bench top rather than under an inert atmosphere, as is often needed for organometallic reactions.

The palladium reagents used are simple to prepare from a range of readily available precursors (aryl halides and trifluoromethanesulfonates). Although the reagents must be synthesized under an inert atmosphere, once made, most can be stored in refrigerated vials for months without degradation, in much the same way as proteins. This might therefore make it possible for kits of various palladium reagents to be made available to researchers who do not have access to the chemicals or equipment needed to make the reagents.

Vinogradova *et al.* went on to show that the organometallic reagents can be used to produce therapeutically interesting compounds known as staple peptides and antibody–drug conjugates (ADCs). In staple peptides, small-molecule linkers bridge two amino-acid residues to reduce the peptide's flexibility. The resulting molecules can have higher affinity for a biological target than non-stapled analogues, or be less susceptible to breakdown by enzymes in the body⁹. The authors used their approach to staple a peptide in less than 10 minutes and with 100% yield.

ADCs⁵ consist of small-molecule drugs attached to antibodies; two ADCs have been clinically approved for treating certain forms of cancer. Such drugs hold the

promise of high efficacy with fewer side effects than small-molecule anticancer drugs because they specifically target the malignant tissue. Clinically approved ADCs are prepared using 'maleimide–thiol' reactions to connect antibodies to the drugs, but these reactions are reversible *in vivo* unless the chemistry of the resulting linkage is carefully designed to prevent this¹⁰. Vinogradova and co-workers used their new reaction to attach the anticancer drug vandetanib to free cysteine residues of an antibody (Fig. 1).

An advantage of the authors' method is that the reaction product is more stable than analogues formed using maleimide–thiol reactions. In tests with one of the products, 83–90% of the peptide remained intact after

treatment with acids, bases and external thiol compounds. Thus, the method might provide a new way to prepare ADCs, as well as other compounds for which stable linkages to proteins are desirable.

The antibodies and staple peptides used in Vinogradova and colleagues' work are fairly stable. It remains to be seen whether more-sensitive proteins that unfold easily can also be used. In addition, the amount of palladium used in the reactions will need to be reduced to low levels, so that unacceptable amounts of the metal are not left as impurities in proteins prepared for medical applications. The authors partly addressed this point by demonstrating that 91–94% of the palladium could be removed from the product. If these two issues can be resolved, then the possibilities for this powerful technology are exciting indeed. ■

Heather Maynard is in the Department of Chemistry and Biochemistry and the California NanoSystems Institute, University of California, Los Angeles, Los Angeles, California 90095, USA. e-mail: maynard@chem.ucla.edu

1. Vinogradova, E. V., Zhang, C., Spokoyny, A. M., Pentelute, B. L. & Buchwald, S. L. *Nature* **526**, 687–691 (2015).
2. Duncan, R. *Nature Rev. Drug Discov.* **2**, 347–360 (2003).
3. Alconcel, S. N. S., Baas, A. S. & Maynard, H. D. *Polym. Chem.* **2**, 1442–1448 (2011).
4. Wu, A. M. *Methods* **65**, 139–147 (2014).
5. Chari, R. V. J., Miller, M. L. & Widdison, W. C. *Angew. Chem. Int. Edn* **53**, 3796–3827 (2014).
6. Fishburn, C. S. J. *Pharm. Sci.* **97**, 4167–4183 (2008).
7. Antos, J. M. & Francis, M. B. *Curr. Opin. Chem. Biol.* **10**, 253–262 (2006).
8. Boutureira, O. & Bernardes, G. J. L. *Chem. Rev.* **115**, 2174–2195 (2015).
9. He, Y., Chen, D. & Zheng, W. *Oncogene* <http://dx.doi.org/10.1038/ncr.2015.37> (2015).
10. Lyon, R. P. *et al. Nature Biotechnol.* **32**, 1059–1062 (2014).

PALAEoANTHROPOLOGY

Homo sapiens in China 80,000 years ago

A discovery in southern China of human teeth dated to more than 80,000 years old indicates that *Homo sapiens* was present in the region considerably earlier than had previously been suspected. [SEE LETTER P.696](#)

ROBIN DENNELL

Debate over when our species, *Homo sapiens*, first dispersed from Africa across southern Asia is hindered by a lack of relevant fossil evidence between the eastern Mediterranean and southeast Asia. Exciting new material is presented in this issue by Liu *et al.*¹ (page 696), who describe a collection of *H. sapiens* teeth from a cave

in southern China's Hunan province. The age and morphology of the teeth suggest that modern humans reached southern China long before they had arrived in northern China or in Europe.

Most researchers agree that our species first appeared in East Africa around 190,000 to 160,000 years ago, and then dispersed into the eastern Mediterranean around 100,000 to 60,000 years ago, after which it was replaced by

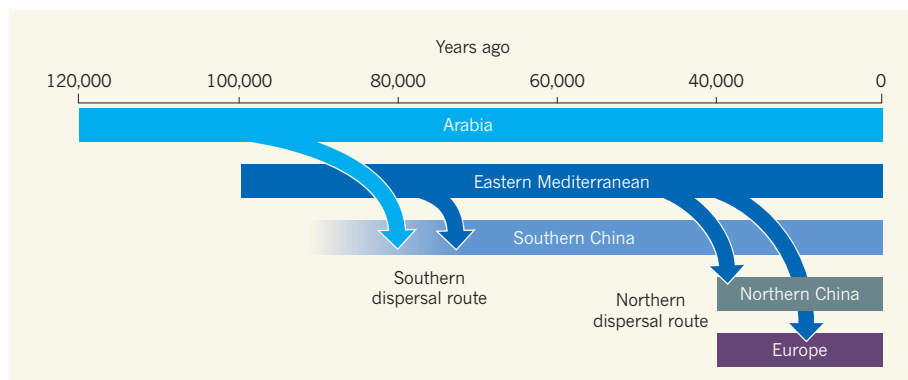


Figure 1 | Human dispersal. Liu and colleagues' discovery¹ of modern-human teeth at Fuyan Cave in southern China implies that *Homo sapiens* dispersed towards the east and south from Arabia or the eastern Mediterranean between 120,000 and 80,000 years ago, considerably earlier than the species is known to have first arrived in Europe and northern China, around 40,000 years ago. The delayed dispersal to these northern regions may be explained by the well-established resident Neanderthal populations in Europe and Siberia, and the harsh, sub-zero winter conditions.

Neanderthals. Following this apparent 'failed dispersal'², one suggested scenario is that *H. sapiens* did not progress eastwards until around 60,000 years ago — a date based on the estimated divergence time of genetic lineages in South Asian people³ and the probable arrival date of humans in Australia⁴. This picture is supported by apparent similarities between small stone tools known as microliths from South Africa dating to around 60,000 years ago and ones from South Asia that are between 30,000 and 36,000 years old^{3,5}. These similarities have been interpreted as showing a direct link between the earliest modern humans in South Asia and their probable African ancestors⁵, although the similarities now seem less robust than was first argued⁶. A contrasting scenario is that *H. sapiens* first dispersed eastwards (probably by way of the Arabian Peninsula) during the last interglacial period, and entered South Asia well before 60,000 years ago, and perhaps before the volcanic supereruption that occurred 74,000 years ago at present-day Lake Toba in Indonesia^{7–9}.

One place where these hypotheses can be tested is in southern China, which is dotted with karst caves that are rich fossil sources. But descriptions of this material have so far been ambiguous because of doubts over the stratigraphic context of skeletal specimens, their dating and/or their identification. Some finds were made by villagers while digging for fertilizer, which obscured the stratigraphic context of the fossils; in other cases, fossil teeth are too worn for identification purposes, or the association between a dated flowstone deposit and a skeletal specimen is unclear¹⁰. At Zhiren Cave in southern China, for example, a mandible (lower jawbone) attributed to *H. sapiens* was found in a geological sequence dated by five flowstones ranging from 110,000 to 55,000 years old¹¹, so the specimen may be younger than the published age of 110,000 years¹²; it has also been suggested that the mandible is from a late *Homo erectus* individual¹³.

Liu and colleagues' discoveries at Fuyan Cave are especially welcome because they seem to lack these usual problems surrounding context, dating and identification. A flowstone (layer 1) covers the entire cave floor, so the underlying material clearly has to be older; a stalagmite on this flowstone was dated to a minimum age of 80,100 years old ($\pm 1,200$ years). The underlying layer 2 is a thin, sandy clay around 20–50 centimetres thick, in which the authors identify numerous mammalian fossils from 38 extant species and 5 extinct large mammals. This fauna is identified as being from the Upper Pleistocene period (125,000 to 10,000 years ago), so the human teeth, which were found in layer 2 in part of the cave, are estimated to be between 80,000 and 120,000 years old.

The teeth are well preserved and show detailed cusp morphology. All 47 are unequivocally attributed to *H. sapiens*. The authors describe them as generally smaller than other African and Asian specimens from the Upper Pleistocene, and closer in size to those of Upper Pleistocene Europeans and contemporary modern humans. The M¹ molar teeth in the sample are different in shape from the rhomboidal contours displayed by Neanderthals or the elongated teeth seen in Asian *H. erectus* fossils. Instead, the relative cusp and occlusal polygon areas of the M¹ molars are almost identical to those of modern Chinese populations. They also seem less primitive than northern Chinese specimens such as those discovered at the Xujiayao site¹⁴.

This discovery has several implications. The finding that *H. sapiens* first appeared in southern China between 120,000 and 80,000 years ago indicates that our species dispersed across southern Asia well before 60,000 years ago (Fig. 1). Furthermore, the fact that the teeth resemble those of Upper Pleistocene Europeans and modern humans implies that the population they came from were immigrants and not the outcome of local evolution from

H. erectus. To place these finds in their continental context, the Fuyan teeth indicate that modern humans were present in southern China 30,000 to 60,000 years earlier than in the eastern Mediterranean and Europe. This is not surprising, perhaps. *H. sapiens* originated in or near the tropics, so it makes sense that the species' initial dispersal was eastwards rather than northwards, where winter temperatures rapidly fell below freezing. As Liu and colleagues point out, the finds also imply that modern humans were in southern China long before there is evidence for them in northern China and Europe. In the case of northern China, the earliest evidence is from Tianyuan Cave near Beijing, dated to around 40,000 years ago¹⁵; in Europe, evidence appears around 45,000 to 40,000 years ago¹⁶. Here, the authors suggest, the presence of Neanderthals may have delayed the arrival of modern humans. However, the predominantly colder winter conditions of the enormous landmass between Europe and northern China may better explain the earlier colonization of southern zones.

Excavation of other caves in the region will undoubtedly add to the findings from Fuyan. What is especially needed now is archaeological evidence (sadly lacking in Fuyan Cave) to indicate whether the initial dispersal of our species was caused or facilitated by cognitive developments (such as symbolism or complex exchange systems), or was simply an example of opportunistic range extension. More revelations about our species' history can surely be expected from southern China. ■

Robin Dennell is in the Department of Archaeology, University of Exeter, Exeter EX4 4QH, UK.
e-mail: r.w.dennell@exeter.ac.uk

1. Liu, W. et al. *Nature* **526**, 696–699 (2015).
2. Shea, J. J. *Quat. Sci. Rev.* **27**, 2253–2270 (2008).
3. Mellars, P., Gori, K. C., Carr, M., Soares, P. A. & Richards, M. B. *Proc. Natl Acad. Sci. USA* **110**, 10699–10704 (2013).
4. Roberts, R. G. et al. *Quat. Sci. Rev.* **13**, 575–583 (1994).
5. Mellars, P. *Science* **313**, 796–800 (2006).
6. Lewis, L., Perera, N. & Petraglia, M. *Quat. Int.* **350**, 7–25 (2014).
7. Dennell, R. & Petraglia, M. D. *Quat. Sci. Rev.* **47**, 15–22 (2012).
8. Boivin, N., Fuller, D. Q., Dennell, R., Allaby, R. & Petraglia, M. D. *Quat. Int.* **300**, 32–47 (2013).
9. Groucutt, H. S. et al. *Evol. Anthropol.* **24**, 149–164 (2015).
10. Dennell, R. W. in *Southern Asia, Australia and the Search for Human Origins* (eds Dennell, R. W. & Porr, M.) 33–50 (Cambridge Univ. Press, 2014).
11. Liu, W. et al. *Proc. Natl Acad. Sci. USA* **107**, 19201–19206 (2010).
12. Kaifu, Y. & Fujita, M. *Quat. Int.* **248**, 2–11 (2012).
13. Dennell, R. W. *Nature* **468**, 512–513 (2010).
14. Xing, S., Martínón-Torres, M., Bermúdez de Castro, J. M., Wu, X. & Liu, W. *Am. J. Phys. Anthropol.* **156**, 224–240 (2015).
15. Shang, H. et al. *Proc. Natl Acad. Sci. USA* **104**, 6573–6578 (2007).
16. Nigst, P. R. et al. *Proc. Natl Acad. Sci. USA* **111**, 14394–14399 (2014).

This article was published online on 14 October 2015.

QUANTUM PHYSICS

Death by experiment for local realism

A fundamental scientific assumption called local realism conflicts with certain predictions of quantum mechanics. Those predictions have now been verified, with none of the loopholes that have compromised earlier tests. [SEE LETTER P.682](#)

HOWARD WISEMAN

The world is made up of real stuff, existing in space and changing only through local interactions — this local-realism hypothesis is about the most intuitive scientific postulate imaginable. But quantum mechanics implies that it is false, as has been known for more than 50 years¹. However, brilliantly successful though quantum mechanics has been, it is still only a theory, and no definitive experiment has disproved the local-realism hypothesis — until now. On page 682 of this issue, Hensen *et al.*² report the first violation of a constraint called a Bell inequality, under conditions that prevent alternative

explanations of the experimental data. Their findings therefore rigorously reject local realism, for the first time.

Bell inequalities are named after John Bell, the physicist who discovered in 1964 that the predictions of quantum mechanics are incompatible with the local-realism hypothesis¹. There are many different ways to make this hypothesis precise³, but Hensen and colleagues' exposition basically follows Bell's original formulation, which states it as the conjunction of two other hypotheses: realism (which Bell called predetermination), essentially meaning that measurements reveal pre-existing physical properties of the world; and locality, roughly meaning that any change

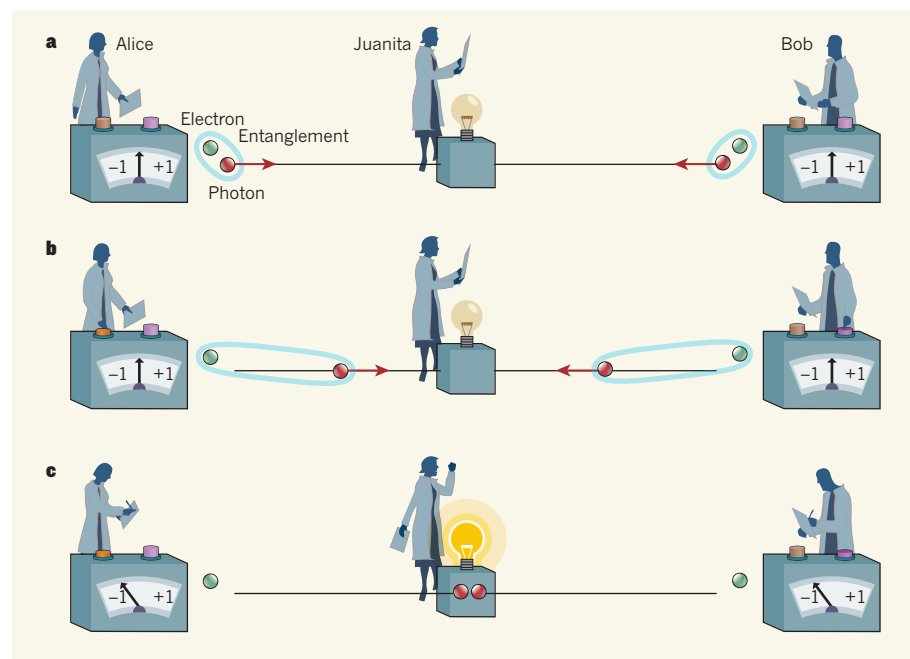


Figure 1 | Violation of a three-party Bell inequality. A Bell inequality is a mathematical relationship regarding the statistics of measurement outcomes obtained by two or more parties. Under certain physical conditions relating to the timing of events, a violation of a Bell inequality proves that local realism — a hypothesis satisfied in all of science except quantum mechanics — is false. Hensen *et al.*² have violated a Bell inequality in such a way that the requisite physical conditions were satisfied for the first time, using the scheme shown in this cartoon. **a**, At separate locations, Alice and Bob create entangled states of an electron and a photon, then send the photons to Juanita's laboratory. **b**, Alice and Bob randomly choose a setting for measurements of their respective electrons. **c**, They obtain their measurement outcomes, and Juanita performs a joint measurement of the photons. Alice's and Bob's outcomes are purely random unless Juanita gets a rare successful outcome (as shown here) that indicates entanglement between Alice's and Bob's electrons. By collating the results over many runs, Hensen *et al.* showed that a Bell inequality had been violated by a statistically significant amount.



50 Years Ago

It may not be generally realized that work is in progress on the colossal project of constructing a 40-in. diameter, 300 miles long, Trans-Alpine oil pipeline to convey oil from the Adriatic to the heart of Germany ... Among the many practical problems concerned with such a project, apart from tunnelling and mechanical excavation in the high Alps, are the necessity to dredge the harbour at Trieste so that it can eventually accommodate oil tankers of 160,000 dead weight tons; setting storage tanks there on piles because available land is a rocky hill site; construction of several thousand feet of piers in the Adriatic ... Involved also in the scheme is the building of five separate pumping stations, each equipped with two 4,000-horsepower electric centrifugal pumps required to lift hundreds of thousands of tons of oil from sea-level to one of the highest points of Felber Tauern.

From *Nature* 30 October 1965

100 Years Ago

'Distances at which sounds of heavy gun-firing are heard' — Referring to the correspondence on this subject, I have been collecting information as to places at which the sound of the firing in Belgium has been heard in this country ... Here, at a distance of about 125 miles from Ypres (taking that town for convenience, as a known centre) I have heard firing quite unmistakably since the beginning of the war — often all day, and for many days in succession, and frequently at night too. So far as I have been able to ascertain, the greatest distance from Ypres at which the firing has been heard unmistakably is about 140 miles ... Observations seem to show that the direction of the wind has less to do with the transmission of the sound than certain atmospheric conditions.

From *Nature* 28 October 1915

enacted at one place cannot have an effect at another place unless there would have been time for a light signal to get from the first place to the second place. The speed of light is relevant because, according to Einstein's theory of relativity, no causal influence can travel faster than light.

A Bell inequality is a mathematical relationship regarding the statistics of measurement outcomes obtained by two or more parties, and also involving the measurement settings chosen by those parties. Suppose that the parties are in well-separated laboratories, and that the measurement settings are chosen and implemented, and the outcomes obtained, in a sufficiently short time that the only way the choice of setting by any party could affect the outcome of any other party would be through a faster-than-light influence. Then, by definition, all Bell inequalities will be satisfied by all local-realistic theories. An experiment violating a Bell inequality therefore implies that either locality or realism is false. Bell's theorem is that, according to quantum mechanics, such an experiment is possible if the parties share particles prepared in a suitable entangled state. Entanglement is a holistic property of a system of quantum particles that can persist even when the particles are far apart.

Bell inequalities have been violated experimentally many times before^{4–9}. However, all of these experiments had loopholes. Either the parties were not far enough apart, given how long it took for the processes involved (randomly choosing a setting, adjusting the apparatus appropriately and obtaining an outcome), or the measurements were inefficient, so that quite often no outcome at all was registered. The inefficiency is relevant because it can allow the existence of local realistic theories — albeit highly contrived ones — that exploit the existence of null outcomes to simulate the correlations of quantum mechanics.

Several groups worldwide have been racing to perform the first Bell experiment that combines large separation, efficient detection and fast operation of the apparatus. Hensen *et al.* have won the race by using a new scheme. Previously, the leading approach was to prepare an entangled state of two photons, send one to one laboratory — conventionally called Alice's — and the other to a second laboratory, Bob's. By contrast, Hensen and colleagues' experiment should be regarded as using a three-party Bell inequality.

In this three-party approach (Fig. 1), Alice and Bob each prepare an entangled state of a photon and an electron, keep their electrons in a diamond lattice and send their photons to Juanita, as I'll call her. Alice and Bob then each choose a setting and measure their electrons, which can be done efficiently, while Juanita performs a joint measurement on the two photons. Alice's and Bob's outcomes will be purely random unless Juanita gets a rare

'successful' result, in which case the outcomes indicate entanglement between Alice's and Bob's electrons. Unlike Alice and Bob, Juanita always makes the same measurement, and so its inefficiency does not open a loophole.

Hensen and co-workers' experiment was made possible only by combining state-of-the-art quantum technologies — it was performed in the Netherlands, but used diamond substrates prepared in the United Kingdom and fast random-number generators developed in Spain. Maintaining optimal operation of all the devices during the experiments was extremely challenging, and the rate of events (defined as Juanita getting a successful outcome) was only about one per hour. As a consequence, only 245 such events were recorded, and the statistical uncertainty in the reported Bell-inequality violation is comparatively large. Nevertheless, from a careful analysis of the entire data set, including runs in which Juanita did not get the desired outcome, Hensen *et al.* reject the local-realism null hypothesis at a confidence level conventionally considered to be statistically significant. It is to be hoped that more data will be generated soon.

The authors' approach might allow them to implement quantum-information protocols enabling secure communication, even when the devices used are not trusted by the users. For this to be practical, the event rate would have to be massively increased above its current level. However, the basic technology and

the scheme (involving joint measurements by the intermediary Juanita) are promising.

The immediate significance of the reported experiment, however, is in hammering the final nail in the coffin of local realism. Some almost metaphysical loopholes remain open — if the results can be replicated with humans, rather than machines, freely choosing the measurement settings and consciously registering the outcomes, then the coffin will have been interred and buried. That experiment, however, is for many years hence. For the moment, we should celebrate Hensen and colleagues' landmark achievement in physics. ■

Howard Wiseman is at the Centre for Quantum Computation and Communication Technology, Centre for Quantum Dynamics, Griffith University, Brisbane QLD 4111, Australia.

e-mail: h.wiseman@griffith.edu.au

1. Bell, J. S. *Physics* **1**, 195–200 (1964).
2. Hensen, B. *et al. Nature* **526**, 682–686 (2015).
3. Wiseman, H. M. *J. Phys. A* **47**, 424001 (2014).
4. Freedman, S. J. & Clauser, J. F. *Phys. Rev. Lett.* **28**, 938–941 (1972).
5. Aspect, A., Dalibard, J. & Roger, G. *Phys. Rev. Lett.* **49**, 1804–1807 (1982).
6. Weihs, G., Jennewein, T., Simon, C., Weinfurter, H. & Zeilinger, A. *Phys. Rev. Lett.* **81**, 5039–5043 (1998).
7. Rowe, M. A. *et al. Nature* **409**, 791–794 (2001).
8. Giustina, M. *et al. Nature* **497**, 227–230 (2013).
9. Christensen, B. G. *et al. Phys. Rev. Lett.* **111**, 130406 (2013).

This article was published online on 21 October 2015.

NON-CODING RNA

Antibiotic tricks a switch

A screen for compounds that block a bacterial biosynthetic pathway has uncovered an antibiotic lead that shuts off pathogen growth by targeting a molecular switch in a regulatory RNA structure. [SEE ARTICLE P.672](#)

THOMAS HERMANN

The golden age of antibiotic discovery, between 1940 and 1960, was heralded by the work of Selman Waksman. A biochemist and microbiologist, Waksman coined the term 'antibiotic' and was the first to use systematic screening to discover antibacterial leads¹. Efforts in his laboratory yielded more than 20 natural antibiotics — most notably streptomycin in 1943, the first effective treatment for tuberculosis². The discovery won Waksman the 1952 Nobel Prize in Physiology or Medicine. Waksman's research caught the attention of scientists at the US pharmaceutical company Merck, and the ensuing collaboration was instrumental in developing

streptomycin for clinical use. On page 672 of this issue, Howe *et al.*³ from the research laboratories of the present Merck describe a new antibiotic lead, identified using a sophisticated refinement of the phenotypic-screening approach introduced by Waksman.

Seven decades after Waksman's research, the flood of antibiotics emerging from natural sources has dwindled to a trickle, and, for various reasons⁴, few companies remain active in antibiotic drug discovery. This is despite an urgent clinical need for new agents in the face of rising resistance to existing antibiotics⁵. On this background, Howe and colleagues' surprising discovery of a drug target in a bacterial non-coding RNA (ncRNA) provides a welcome bright spot.

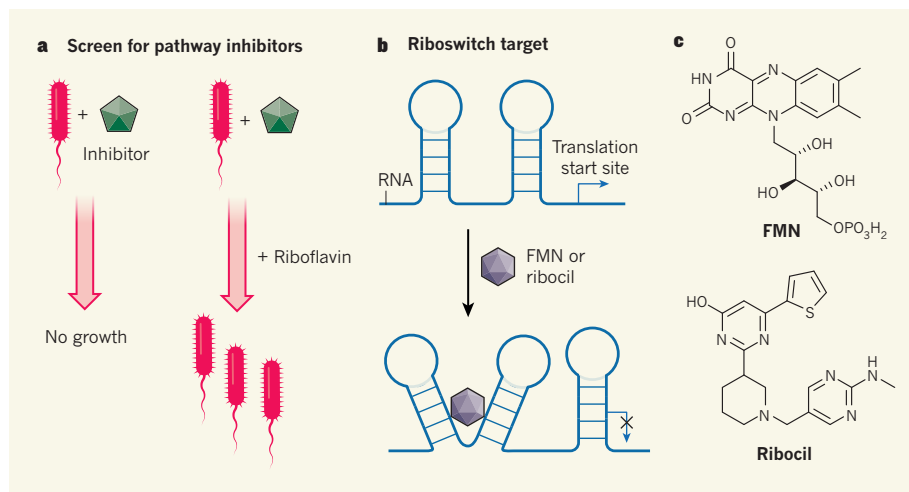


Figure 1 | Metabolic-pathway blockade identifies an antibiotic. **a**, Bacteria use a metabolic pathway to synthesize riboflavin — a molecule involved in many enzymatic reactions — when environmental riboflavin is not available. Howe *et al.*³ searched for inhibitors of this pathway by identifying compounds whose antibacterial action could be attenuated by riboflavin addition. **b**, The screen uncovered the synthetic molecule ribocil as a selective inhibitor of riboflavin biosynthesis. The authors report that ribocil binds to a ‘riboswitch’ regulatory domain in a non-coding region of the messenger RNA that encodes a synthase enzyme involved in riboflavin synthesis. The riboswitch is normally bound by flavin mononucleotide (FMN), a metabolite produced from riboflavin; such binding induces a structural change in the riboswitch that prevents expression of the synthase enzyme and thus stops further riboflavin synthesis when sufficient FMN is available. **c**, Ribocil, despite having a very different structure from FMN, also binds at this site, tricking the switch into shutting off riboflavin production and depriving the bacterium of this essential metabolite.

Phenotypic screening involves applying test compounds to bacterial cultures and observing changes to the cells’ characteristics (their phenotype). Since Waksman’s days, the simplest phenotypic screen for antibacterial activity has been to look for growth inhibition — cell death is the most severe of phenotypes and is readily measured. Howe and colleagues established a more subtle screen that interrogates a single bacterial metabolic pathway. Unlike humans, bacteria can synthesize the metabolite riboflavin, also called vitamin B2, which is a precursor of the cofactors required in many enzymatic reactions. The most prominent of these cofactors is flavin mononucleotide (FMN). Genes involved in riboflavin biosynthesis are essential only when the bacteria cannot acquire the vitamin from its environment. The researchers hypothesized that antibacterial compounds whose growth-suppressing effect could be reversed by supplementing the bacterial cultures with riboflavin would be candidate inhibitors of the riboflavin biosynthesis pathway (Fig. 1a). They tested around 57,000 synthetic small molecules and identified the molecule ribocil as one that kills bacteria by selectively blocking riboflavin biosynthesis.

To confirm ribocil’s mechanism of action, Howe and colleagues demonstrated that bacteria treated with the compound were indeed depleted in riboflavin. In mice infected with pathogenic bacteria, ribocil treatment reduced the concentration of bacteria by more than 1,000-fold, which is a promising start. The team went on to identify the molecular target of ribocil by isolating bacterial clones that

became resistant during prolonged exposure to ribocil at sublethal concentrations, and sequencing their genomes. What initially seemed a routine exercise turned to excitement when the resistant bacteria were found to harbour mutations in a non-coding stretch of the bacterial genome, suggesting that ribocil blocks a mechanism of gene regulation rather than inhibiting a protein target. Further sleuthing for ribocil’s site of attack led the researchers to a highly structured ncRNA domain upstream of the sequence that marks the translational start site in a messenger RNA encoding a key synthase enzyme in the riboflavin biosynthesis pathway.

It turns out that the ncRNA domain bound by ribocil is normally bound by FMN, and the site is a ‘riboswitch’, the term used for RNA regions that change structure when bound by a ligand (Fig. 1b). Binding of FMN traps the riboswitch RNA in a configuration that prevents access of transcription and translation machinery and thus blocks expression of the synthase enzyme. This mechanism provides a way for bacteria to reduce further production of riboflavin, and thus FMN, when sufficient amounts of the cofactor are available. By binding to the same ncRNA target, ribocil tricks the riboswitch to respond, shutting off riboflavin production and depriving the bacteria of the essential metabolite. As a compelling piece of evidence, the researchers used X-ray crystallography to provide a snapshot of ribocil in the act of binding to the riboswitch RNA. Such structural information will be valuable for improving the antibacterial activity of ribocil

derivatives for potential clinical use.

This mechanism of mimicking a natural ligand of a riboswitch is not known from any other antibacterial compounds. The study thus conclusively demonstrates the advantages of phenotypic screening that allows all constituent components of a pathway — proteins and nucleic acids — to be tested simultaneously in an unbiased fashion. Although riboswitches are obvious targets for antibacterial discovery⁶, Howe and colleagues have for the first time identified a riboswitch-binding molecule that is not a close structural analogue of a metabolite ligand. This improves the odds that ribocil will have no off-target effects on other pathways that involve riboflavin and FMN, which is already hinted at by the authors’ observation that even high doses of the compound were not toxic in mice.

From a broader perspective, Howe and colleagues’ research is a striking demonstration that structured regions in ncRNA may serve as targets for synthetic drugs. There are only a few previous instances of ncRNA elements being exploited as potential drug targets, one being a region in the hepatitis C virus that is bound by synthetic inhibitors of viral protein synthesis⁷. But many natural antibiotics⁸ work by interfering with protein synthesis through targeting ncRNA components in the bacterial ribosome, the cellular machine that synthesizes all proteins.

Assuming that nature is on to something by targeting ncRNA in the ribosome, an optimist might point out that the plethora of ncRNAs recently discovered in the genomes of bacteria and more-complex organisms will provide a rich expansion of the range of targets for therapeutic intervention^{9,10}. For the sceptics, Howe and colleagues’ work might allay doubts that targeting ncRNA elements may deliver new drugs. And we can hope that clinicians who urgently need alternative antibiotics will be among the first to reap the benefits of RNA emerging as a drug target. ■

Thomas Hermann is in the Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, California 92093, USA.
e-mail: tch@ucsd.edu

1. Waksman, S. A. *Mycologia* **39**, 565–569 (1947).
2. Waksman, S. A. *Science* **118**, 259–266 (1953).
3. Howe, J. A. *et al. Nature* **526**, 672–677 (2015).
4. Infectious Diseases Society of America. *Bad Bugs, No Drugs* (IDSA, 2004).
5. World Health Organization. *Antimicrobial Resistance: Global Report on Surveillance* (WHO, 2014).
6. Matzner, D. & Mayer, G. *J. Med. Chem.* **58**, 3275–3286 (2015).
7. Dibrov, S. M. *et al. J. Med. Chem.* **57**, 1694–1707 (2014).
8. McCoy, L. S., Xie, Y. & Tor, Y. *Wiley Interdisc. Rev. RNA* **2**, 209–232 (2011).
9. Morris, K. V. & Mattick, J. S. *Nature Rev. Genet.* **15**, 423–437 (2014).
10. Taft, R. J., Pang, K. C., Mercer, T. R., Dinger, M. & Mattick, J. S. *J. Pathol.* **220**, 126–139 (2010).

This article was published online on 30 September 2015.

Projections from neocortex mediate top-down control of memory retrieval

Priyamvada Rajasethupathy^{1,2*}, Sethuraman Sankaran^{2*}, James H. Marshel^{1*}, Christina K. Kim^{1,3}, Emily Ferenczi^{1,3}, Soo Yeun Lee^{1,3}, Andre Berndt^{1,3}, Charu Ramakrishnan¹, Anna Jaffe¹, Maisie Lo¹, Conor Liston^{1,4} & Karl Deisseroth^{1,2,4,5}

Top-down prefrontal cortex inputs to the hippocampus have been hypothesized to be important in memory consolidation, retrieval, and the pathophysiology of major psychiatric diseases; however, no such direct projections have been identified and functionally described. Here we report the discovery of a monosynaptic prefrontal cortex (predominantly anterior cingulate) to hippocampus (CA3 to CA1 region) projection in mice, and find that optogenetic manipulation of this projection (here termed AC-CA) is capable of eliciting contextual memory retrieval. To explore the network mechanisms of this process, we developed and applied tools to observe cellular-resolution neural activity in the hippocampus while stimulating AC-CA projections during memory retrieval in mice behaving in virtual-reality environments. Using this approach, we found that learning drives the emergence of a sparse class of neurons in CA2/CA3 that are highly correlated with the local network and that lead synchronous population activity events; these neurons are then preferentially recruited by the AC-CA projection during memory retrieval. These findings reveal a sparsely implemented memory retrieval mechanism in the hippocampus that operates via direct top-down prefrontal input, with implications for the patterning and storage of salient memory representations.

Pioneering studies (recently reviewed¹) have illuminated the molecular and physiological mechanisms of information storage at synapses, but how populations of individual neurons form network representations of memory is largely unknown. Recent studies have reported that only a fraction of eligible neurons are allocated during learning to form a memory^{2–4}, that the population initially encoding the memory is later preferentially recruited during retrieval of that memory⁵, and that subsequent activation of the initial ensemble alone can be sufficient for memory retrieval^{6–8}. Intriguing questions remain regarding (for instance) whether neurons in the memory ensemble have different roles in storage and retrieval, or are controlled by top-down influences^{9–11} distinct from the well-explored bottom-up entorhinal/hippocampal system; such top-down prefrontal projections to the hippocampus (if functionally present) might also support bidirectional communication during memory consolidation¹², and would potentially be relevant to psychiatric disorders such as post-traumatic stress disorder¹³, schizophrenia¹⁴ and drug addiction¹⁵. It is also unclear whether behaviourally salient memories are laid down broadly across the brain^{16,17}, or are wired topographically within the local network for improved access^{18–21}. To address these and other questions regarding real-time population-level mechanisms of memory storage and retrieval, we developed an approach to visualize and perturb top-down modulation of rapidly evolving memory ensembles in behaving mammals.

AC-CA: a direct top-down projection

To identify direct top-down inputs to the hippocampus, we injected a retrograde tracer capable of labelling afferent neurons with tdTomato (RV-tdT²²) into the hippocampus. We observed robust tdT labelling in brain regions with known inputs to the hippocampus, including the medial septum, contralateral CA3 and entorhinal cortex (Fig. 1a and Extended Data Fig. 1a). Additionally, we identified a previously

uncharacterized input arising from the dorsal anterior cingulate cortex (AC) and adjacent frontal cortical association cortex, both of which are reciprocally connected with the mediodorsal thalamic nucleus—a defining feature of the prefrontal cortex (PFC) in rodents (Fig. 1a; also confirmed with another retrograde tracer, canine adenovirus (CAV)²³; Extended Data Fig. 1b). Injection of RV-tdT in the AC also sparsely labelled neurons bilaterally in the dorsal hippocampus, consistent with potential bidirectional communication between the AC and hippocampus (Extended Data Fig. 1c). To validate further the existence of this novel PFC-to-hippocampus projection, we injected an anterograde label (adeno-associated virus 5-enhanced yellow fluorescent protein (AAV5-eYFP)) into the dorsal AC (Fig. 1b) and detected fluorescence-filled projection terminals bilaterally in the striatum and ipsilaterally in the medial dorsal thalamic nucleus (both areas are known to receive projections from the PFC), but also bilaterally in the hippocampus.

To determine if these prefrontal projections gave rise to direct monosynaptic drive of hippocampal neurons, we transduced the AC with an AAV encoding a channelrhodopsin (ChR), and performed patch-clamp recordings of light-driven excitatory postsynaptic currents (EPSCs) in CA1/CA3 cell bodies (Fig. 1c). Cells in both CA1 (Fig. 1d) and CA3 (Fig. 1f) reliably responded to light pulse trains, and generated evoked EPSC amplitudes sufficient to drive action potentials (Fig. 1h, i). Responses when present were fast, with mean latency of 3.2 ms in CA1 ($n = 26$; Fig. 1e) and 2.7 ms in CA3 ($n = 13$; Fig. 1g); along with the observation of sustained evoked spikes after 10 Hz stimulation (Fig. 1h, i), this finding was consistent with the presence of a direct and efficacious monosynaptic connection from the AC onto hippocampal pyramidal cells in the CA3/CA1 subfields, which we accordingly term the AC-CA projection. No responses were observed in dentate neurons (Fig. 1j).

¹Department of Bioengineering, Stanford University, Stanford, California 94305, USA. ²CNC Program, Stanford University, Stanford, California 94305, USA. ³Neuroscience Program, Stanford University, Stanford, California 94305, USA. ⁴Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, California 94305, USA. ⁵Howard Hughes Medical Institute, Stanford University, Stanford, California 94305, USA.

*These authors contributed equally to this work.

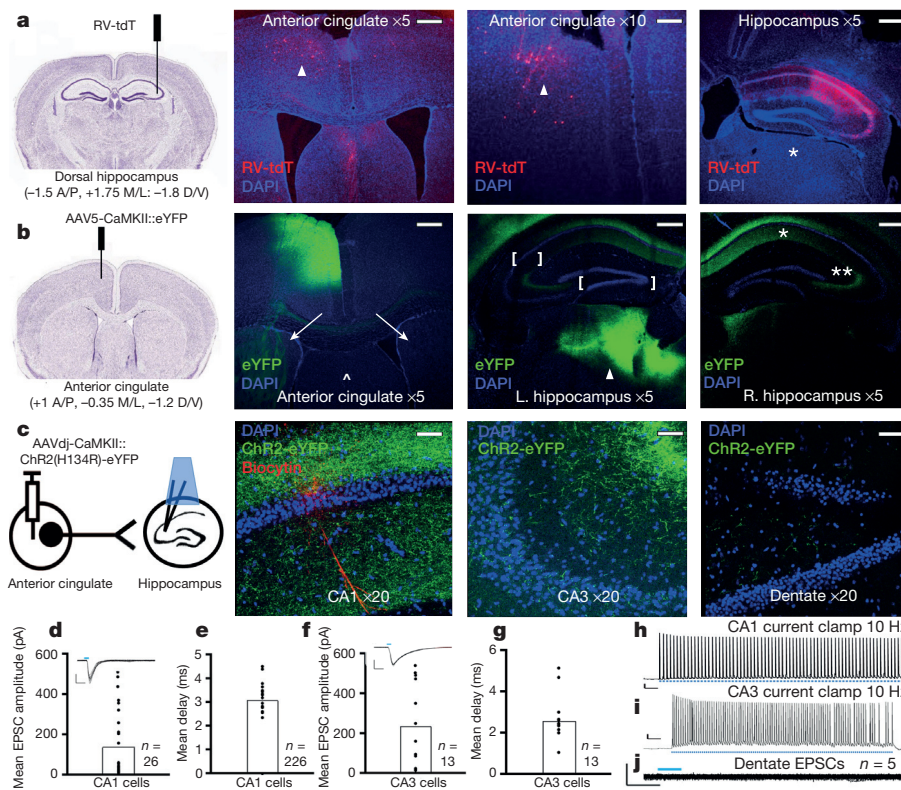


Figure 1 | Characterization of AC-CA monosynaptic projection. **a**, Five days after RV-tdT injection into the hippocampus (coordinates specified), retrogradely labelled neurons were detected in the AC (arrowhead). The injection did not leak below the hippocampus into the medial dorsal thalamus (asterisk), a known projection target of the PFC. Scale bars: $\times 5$: 300 μm ; $\times 10$: 100 μm (confocal). A, anterior; D, dorsal; DAPI, 4',6-diamidino-2-phenylindole; M, medial; L, lateral; V, ventral. **b**, Five weeks after AAV5-CaMKII::eYFP injection into the AC (coordinates specified), projection fibres were visualized in the bilateral striatum (arrows), the bilateral hippocampus at the stratum oriens and stratum radiatum of CA1 (one asterisk) and CA3 (two asterisks), and the ipsilateral medial dorsal thalamus (arrowhead), while

AC-CA: causal role in memory retrieval

To probe the functional importance of this projection, we conducted a series of optogenetic experiments to manipulate this pathway before and after contextual fear conditioning, and also in the setting of memory extinction and reinstatement. We first injected RV-ChR2-eYFP (or RV-eYFP) into the dorsal hippocampus, and targeted light delivery to retrogradely labelled cell bodies in the AC (Extended Data Fig. 2a). On day 1, ChR2 and eYFP mice underwent contextual fear conditioning in one context, while a ChR2-expressing control group was exposed to the same context without shock. On day 2, all mice were placed in a different context, in which the ChR2-expressing fear-conditioned group ($n = 8$) showed significant fear behaviour (freezing) only during light stimulation, compared with unconditioned ChR2 (no shock, $n = 6$) or shocked control groups (eYFP, $n = 6$, $P < 0.001$, two-way analysis of variance (ANOVA) with repeated measures; Extended Data Fig. 2b). The time to freezing and time to unfreezing with light on/off switching were largely consistent across animals. On day 3, mice were placed back in the original context, verifying in both ChR2 and eYFP fear-conditioned cohorts strong memory encoding and retrieval, with significantly greater levels of freezing compared with no-shock controls (Extended Data Fig. 2c; $P < 0.001$, unpaired t -test).

This observation that cells contributing to the AC-CA projection can activate contextually conditioned fear behaviour was replicated and extended using a complementary anterograde projection-targeting strategy. We injected AAVdj-ChR2-eYFP (or AAVdj-eYFP

sparing CA2 and the dentate (brackets). The injection did not leak into the medial septum, a known input to the hippocampus (caret). Scale bars: $\times 5$: 300 μm (100 μm maximum projections). L., left; R., right. **c**, AAVdj-CaMKII::ChR2(H134R)-eYFP was injected into the dorsal AC; post-synaptic responses were recorded from CA1 pyramidal cells in acute slice. Scale bars: $\times 20$: 60 μm . **d**, CA1 response amplitudes ($n = 26$ neurons, $n = 6$ mice); inset: raw traces. **e**, CA1 response latency (mean 3.2 ms). **f**, CA3 response amplitudes ($n = 13$ neurons, $n = 2$ mice); inset: raw traces. **g**, CA3 response latency (mean 2.7 ms). **h**, **i**, CA1 (**h**) and CA3 (**i**) current-clamp traces illustrating spiking reliability at 10 Hz. Scale: 10 mV, 250/500 ms. **j**, No responses were detected in dentate neurons ($n = 5$).

in a parallel cohort) into the AC, and targeted light stimulation to terminals in the hippocampus (Fig. 2a). We again observed significant freezing to optical stimulation in the neutral context only in the ChR2 group (Supplementary Video 1) compared with the no-shock and eYFP controls (Fig. 2b; $n = 8$ for all groups; $P < 0.001$, two-way ANOVA with repeated measures). These animals exhibited the same characteristic latency to freezing during light stimulation as in the previous experiment. We next tested whether this consistent behavioural response was indeed due to reactivation of a fear memory, rather than due to direct nonspecific drive of fear behaviour. The same mice corresponding to those shown in Fig. 2a, b were subjected to several days of contingency degradation by exposure to context A without shock (Methods), after which stimulation with light failed to produce significant freezing in ChR2 animals, as with no-shock and eYFP controls (Fig. 2c and Supplementary Video 2). Fear conditioning was then reinstated in these mice in a new context, after which light stimulation once again reliably produced freezing in ChR2 mice compared with no-shock and eYFP controls (Fig. 2d; $n = 8$ for all groups; $P < 0.001$, two-way ANOVA with repeated measures; Supplementary Video 3). Preservation of contextual fear memory on day 3 and successful fear memory extinction on day 14 were confirmed (Fig. 2e). While these data demonstrated that cells contributing to the AC-CA projection can drive fear memory recall, it remained possible that any drive of the hippocampus could induce retrieval of a recent strongly represented memory in the hippocampus. However, we found no evidence for this possibility

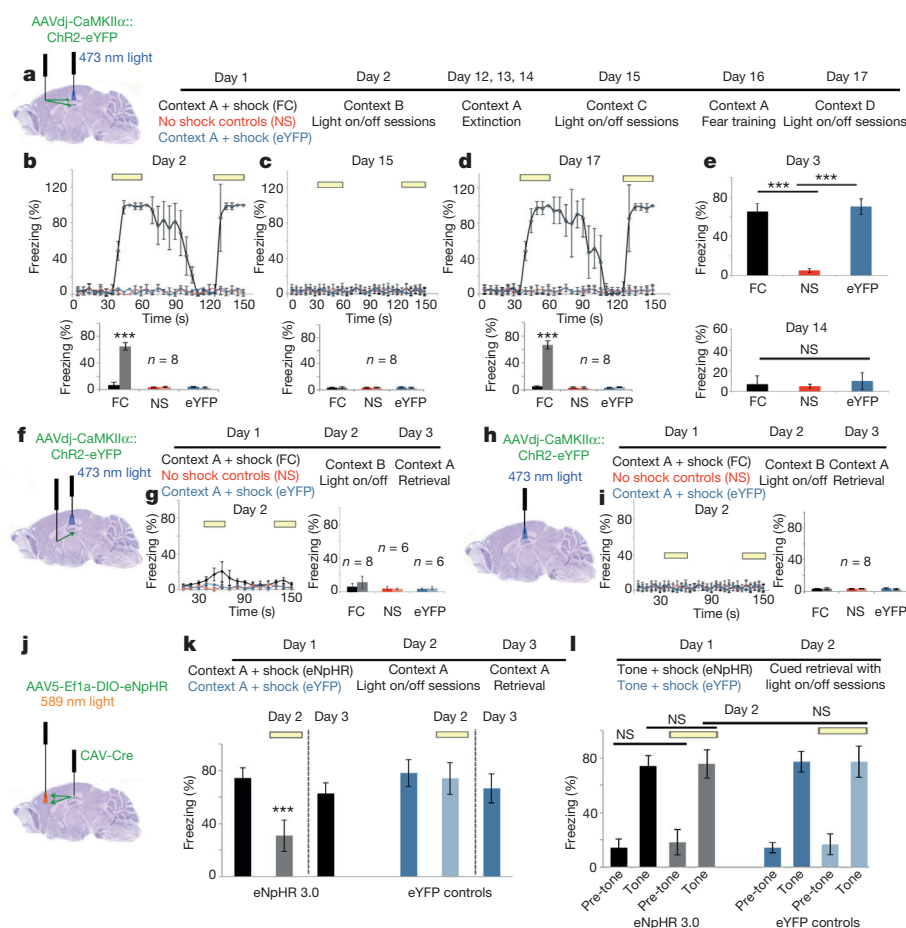


Figure 2 | AC-CA projections control top-down memory retrieval.

a, AAVdj-CaMKII α ::ChR2-eYFP (or eYFP alone) was injected into the AC, and light was targeted to the dorsal hippocampus 5 weeks after injection. Timelines are indicated. FC, fear context; NS, no shock. **b–d**, Percentage time freezing on days 2, 15 and 17: 5 s time bins ($n = 8$; $P < 0.001$, two-way ANOVA with repeated measures) are shown. Throughout this figure, error bars represent standard deviation (s.d.), to facilitate comparison with previously published literature; however, since freezing behaviour is not unbounded, these error bars may overestimate the upper limit of error. Summary bar graphs below are mean \pm s.d. 20 s before light (dark bar) versus 20 s after light (lighter bar). **e**, Preservation of contextual memory (day 3) and contextual extinction (day 14), original context A (mean \pm s.d., $n = 8$, $P < 0.001$, paired t -test). **f**, AAVdj-CaMKII α ::ChR2-eYFP (or eYFP alone) was injected into the medial septum, and light stimulation was targeted to the dorsal hippocampus 5 weeks after injection. **g**, Percentage time freezing during

with two additional control experiments, either directly driving a different (septo-hippocampal) projection (Fig. 2f, g) or directly driving the hippocampus itself (Fig. 2h, i). Preservation of normal contextual fear memory in these mice was confirmed as before (Extended Data Fig. 2d, e; $n = 8$, $P < 0.001$, paired t -test).

These experiments demonstrated that activating AC-CA projecting cells was sufficient to induce contextual memory retrieval; we next tested necessity by targeting the inhibitory opsin eNpHR3.0 to cells giving rise to the projection (Extended Data Fig. 2f, g), with light targeted focally and bilaterally to AC-CA cell bodies (Fig. 2j). We observed striking deficits in both the latency (Extended Data Fig. 2h) and the strength (Fig. 2k; $n = 12$ for eNpHR3.0 group, $n = 8$ for eYFP group, $P < 0.001$, two-way ANOVA with repeated measures) of the fear response in the eNpHR3.0 group (Supplementary Video 4) compared with eYFP controls (Supplementary Video 5). This effect was fully reversible (Fig. 2k). We also found that eNpHR3.0 mice demonstrated intact auditory cued memory recall (Fig. 2l; $n = 8$ for all groups, not significant with paired t -tests), confirming that the loss-

of-function experiments described earlier represented a hippocampus-specific effect of the AC-CA projecting cells. Taken together, these anatomical, electrophysiological and behavioural data reveal the existence of a previously uncharacterized monosynaptic PFC-to-hippocampus projection. When this circuit is inhibited, fear-conditioned mice are unable to retrieve the fear memory with the same strength or speed as control counterparts, indicating endogenous importance for memory retrieval. In contrast, activation of this circuit is sufficient for robust fear memory retrieval in recently conditioned mice, but not in naive unconditioned mice, mice in which the memory has been extinguished, mice not expressing opsin, or mice receiving other types of direct or indirect drive of hippocampus.

of-function experiments described earlier represented a hippocampus-specific effect of the AC-CA projecting cells.

Taken together, these anatomical, electrophysiological and behavioural data reveal the existence of a previously uncharacterized monosynaptic PFC-to-hippocampus projection. When this circuit is inhibited, fear-conditioned mice are unable to retrieve the fear memory with the same strength or speed as control counterparts, indicating endogenous importance for memory retrieval. In contrast, activation of this circuit is sufficient for robust fear memory retrieval in recently conditioned mice, but not in naive unconditioned mice, mice in which the memory has been extinguished, mice not expressing opsin, or mice receiving other types of direct or indirect drive of hippocampus.

Highly correlated neurons emerge during learning

To observe real-time influences of the AC-CA projection on hippocampal network activity, we adapted the fear-conditioning paradigm to head-fixed mice navigating in a virtual environment on an axially

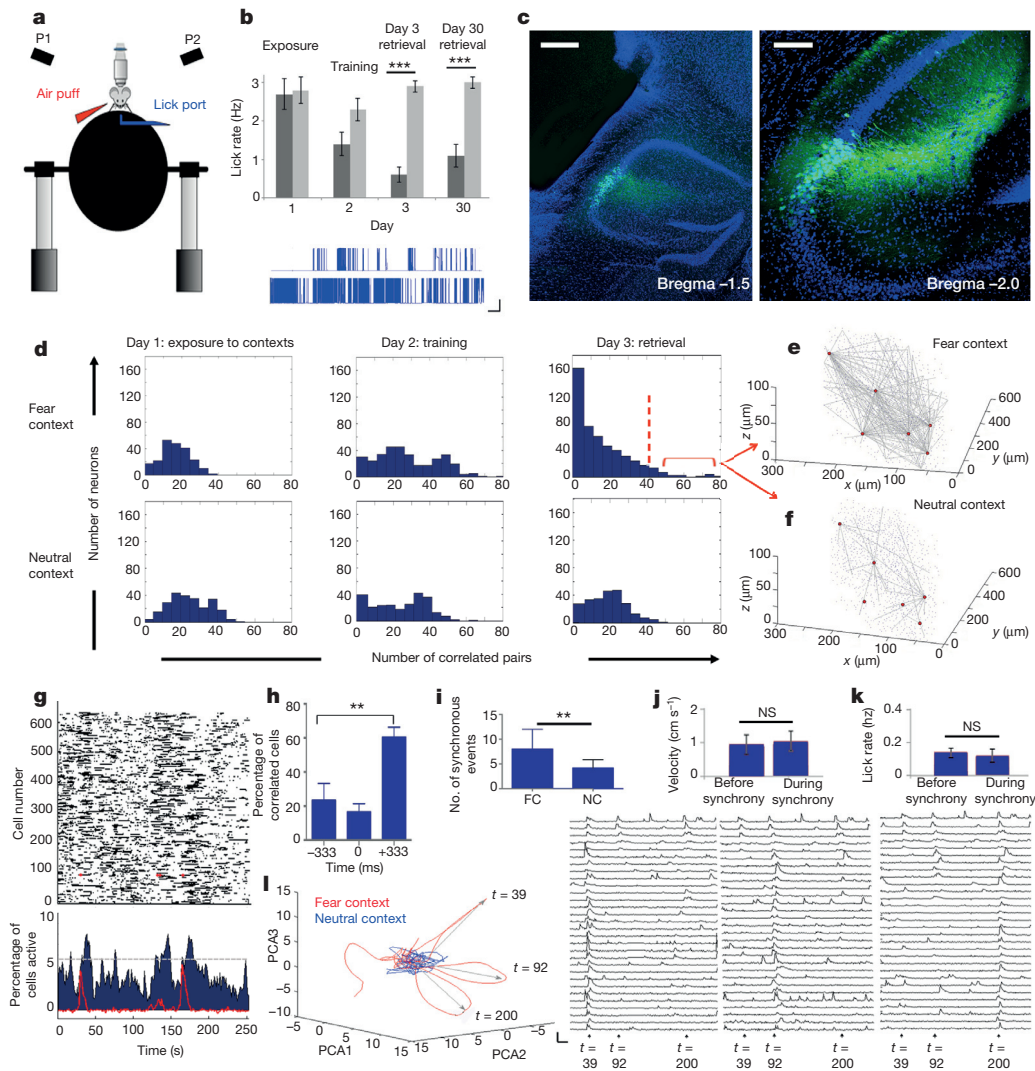


Figure 3 | Memory formation generates highly correlated HC neurons that represent context. **a**, Head-fixed virtual reality setup. Fear was quantified by lick suppression (Methods). P1, projector 1; P2, projector 2. **b**, Lick rates in fear (black) versus neutral (grey) contexts (mean \pm s.d.; $n = 12$; $P < 0.001$, paired t -test). Sample day 3 raw lick profiles in the fear (top) versus neutral (bottom) context. Scale: 1 V, 20 s. **c**, Histology performed 4 weeks after injection/surgical implantation showing implant above GCaMP6m-expressing neurons. Left: $\times 10$; scale bar, 200 μm . Right: $\times 20$; scale bar, 50 μm . **d**, Histograms showing number of correlated partners for each neuron in the fear versus neutral contexts for a representative mouse (see Extended Data Fig. 5). **e**, **f**, HC neurons in the fear context (red) (**e**) have few correlated partners in the neutral context (**f**). $n = 4$ mice (60 ± 18.2 (s.d.) in the fear context versus 18 ± 15.8 in the neutral context; $P < 0.01$, paired t -test; see Extended Data Fig. 6). **g**, Raster plot (top) and collapsed activity histogram (bottom) during

fixed track ball under a two-photon microscope²⁴. We used lick suppression, rather than immobility, as a measure of fear behaviour in water-restricted mice^{25–27} (Fig. 3a and Methods); mice learned this task and displayed significant lick suppression during retrieval in the fear context, indicating successful memory retrieval (Fig. 3b; $n = 12$, $P < 0.01$, paired t -test). For imaging during behaviour, mice were injected with the genetically encoded Ca^{2+} indicator GCaMP6m²⁸, implanted with a cranial window above CA2/CA3 (Fig. 3c; confirmation of normal hippocampal physiology and behaviour in these mice is shown in Extended Data Fig. 3), and imaged daily in both contexts during training and retrieval (Extended Data Fig. 4a). We targeted imaging preferentially to CA3 neurons by both stereotactic positioning of the imaging field of view (FOV) (Fig. 3c), and by verifying that visualized neurons displayed dendritic processes traveling in the

memory retrieval in one mouse; representative HC neuron time series is overlaid (red). **h**, HC neuron activity onset (time 0) compared with onset activity of their correlated pairs ($n = 67$ HC neurons, $60.3 \pm 6\%$ leading versus $23.2 \pm 10\%$ lagging; $P < 0.01$, unpaired t -test). **i–k** Synchronous activity (Methods) was quantified ($n = 5$, 8.1 ± 4 events in the fear versus 4.2 ± 1.6 in the neutral context; $P < 0.01$, paired t -test) (**i**), and was not accompanied by significant changes in velocity (**j**) or lick rate (**k**) ($n = 5$ mice; not significant (NS), paired t -test). FC, fear context; NC, neutral context. **l**, Principal component analysis (PCA) from a representative mouse (Extended Data Fig. 9 for additional data sets). Population trajectory in the fear (red) versus the neutral context (blue) was projected onto the respective first three principal components. Right: $\Delta F/F$ traces of HC neurons and their correlated neurons participating in each deflection. F , fluorescence; Scale: $400\% \Delta F/F$, 20 s. All error bars represent s.d. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.

focal plane of this dorsal view (Supplementary Video 6; in contrast, CA2 or CA1 neurons show processes traveling out of the focal plane due to the curvature of the pyramidal layer); however, some CA2 neurons are likely to be included. In all cases, fast volumetric ($500 \times 500 \mu\text{m} \times y$, $100 \mu\text{m} z$) two-photon imaging was performed, providing access to >400 neurons (Extended Data Fig. 4b).

We first sought to identify features of the functional network that differed consistently across days or contexts. Many features were indistinguishable, including mean response magnitude of active neurons, mean activity event rate, mean activity event duration, and spatial distribution of active neurons (Extended Data Figs 4c–f and 5a). However, we noted a significant increase in the number of neurons active during memory retrieval in the fear context (Extended Data Fig. 5b; $n = 5$ mice); interestingly, this was accompanied by

a significant decrease in mean correlated activity. Although prior studies reported increased correlated activity after learning^{29–31}, a parsimonious unifying explanation could be that reduced mean correlated network activity reflects a state of greater sparsity after learning, in which few neurons engage in considerably higher correlated activity while most become de-correlated. Indeed, binning the number of correlated partners for each neuron revealed a significant distribution shift (Fig. 3d and Extended Data Fig. 5d), from randomly organized Poisson-like correlation distributions before learning to more ordered, power-law-like distributions after learning, with the emergence of a small population of neurons displaying highly correlated activity in the network (HC neurons). Importantly, these network properties were specific to the learned context, and indeed, at the single-neuron level, the highest levels of cell-by-cell correlation were seen in the fear context compared to the neutral context (Fig. 3d and Extended Data Fig. 5d; $n = 5$; $P < 0.01$, paired t -test). Additional quantitative properties were assessed for differential representation of the fear and neutral contexts (Extended Data Fig. 5d–i), revealing that the power-law exponent b provided the largest contribution to this context separation (Extended Data Fig. 5c; $n = 5$, $P < 0.01$), consistent with the emergence of HC neurons representing aspects of fear memory retrieval. Intriguingly, HC neurons in the fear context (Fig. 3e, shown in red) tended to be neurons that had a low degree of correlated partners in the neutral context (Fig. 3f and Extended Data Fig. 6), suggesting that the emergence of HC neurons after learning does not stem simply from strengthening of pre-existing correlated cell assemblies.

To understand better the importance of these HC neurons, we next focused analysis on the activity of the entire network at times when the HC neurons were active. The HC neurons tended to lead rather than lag their correlated pairs (Fig. 3h), which were spatially distributed

throughout the volume (Extended Data Fig. 7a, b). Furthermore, although overall cell-by-cell correlated activity was reduced during the fear retrieval test, significantly more population-wide synchronous events (Fig. 3g, i), which were confirmed to be not related to motion (Fig. 3j, k) and consisted of essentially orthogonal groups of neurons (Fig. 3l and Extended Data Fig. 9), occurred in the fear context; HC neurons were found to lead these broad synchronous events (Extended Data Fig. 7c; $P < 0.001$, Kolmogorov–Smirnov two-tail test), with 78% of HC neurons active within the first 20% of a synchronous event. This event-leading nature could be consistent with a role for HC neurons in recruiting network activity.

Importantly, these analyses were designed to limit the effects of potential confounds of slow fluctuations in the signal (for example, GCaMP6m and Ca^{2+} kinetics) on correlations between neurons (see Methods). Additionally, a fast non-negative deconvolution analysis³² (detailed fully in Methods), for detecting onset of activity while removing slow decay kinetics, yielded consistent results as described earlier for the increased pair-wise correlations in the fear versus neutral context, the event-leading nature of HC neurons compared with their correlated pairs, and the increased synchronous events observed in the fear versus neutral context (Extended Data Fig. 8).

AC–CA projections target HC neurons during retrieval

These volumetric imaging studies during memory retrieval demonstrate the emergence of a sparse set of HC neurons characterized by high correlations and leading of local synchronous events. Such neurons could serve as efficient points of access if preferentially recruited by top-down projections during memory retrieval. To test this idea, we sought to stimulate AC–CA projections while simultaneously imaging the postsynaptic hippocampal network to observe local dynamics directly. Current *in vivo*-tested red-shifted

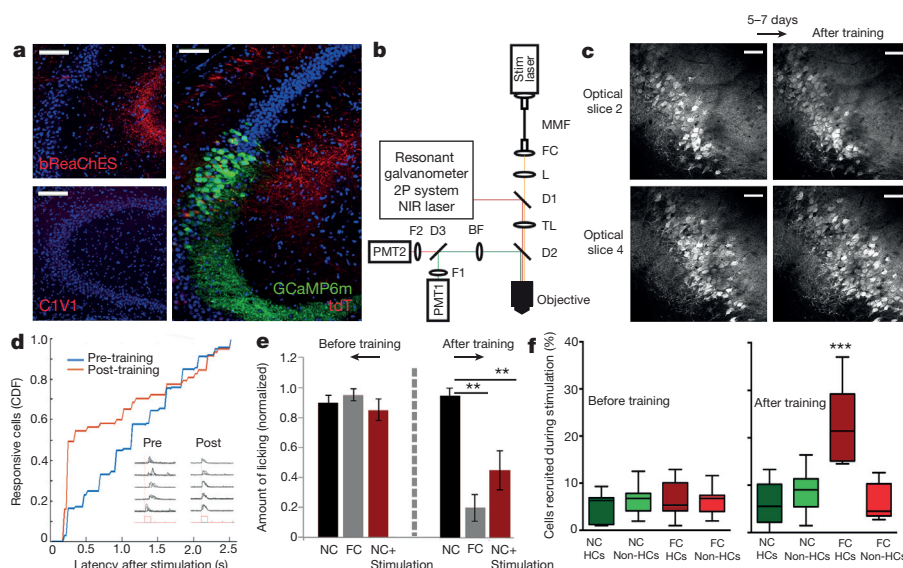


Figure 4 | The AC–CA projection preferentially recruits HC neurons during memory retrieval. **a**, Left, AAV8–CaMKII::bReaChES–eYFP or AAV8–CaMKII::C1V1_{TT}–eYFP (C1V1_{TT} denotes the red-light-activated channel-rhodopsin engineered and characterized in refs 33 and 47) was injected into the AC, and fibre terminals were visualized in CA3 (red). Scale bars, 60 μm ($\times 20$). Right, AAV8–CaMKII::tdT in the AC, and AAVdj–CaMKII::GCaMP6m in CA2/CA3. Histology 8 weeks after cannula implantation showing preservation of AC–CA projections (red) near GCaMP6m-expressing neurons (green). Scale bar, 50 μm ($\times 20$). **b**, Two-photon imaging and full-field optogenetic stimulation setup (Methods). Green indicates GCaMP6m signals; red, stimulation artefact. D1, 680 nm short-pass dichroic; D2, 594 band/NIR long-pass dichroic; D3, 555 nm long-pass dichroic; FC, fibre couple; F1, 520 \pm 22 nm filter; F2, 624 \pm 20 nm filter; L, 25 mm planoconvex lens; MMF, multimode fibre; NIR BF, near infrared blocking filter; PMT1 and PMT2, GaAsP PMTs; TL, tube lens.

c, Z-projection images (mean over time) at two depths (40 μm apart) from a representative mouse before and after training. Scale bars, 60 μm . **d**, Proportion of cells responding (CDF, cumulative distribution function) to AC–CA stimulation as a function of latency ($n = 4$ mice; 12 trials, $P = 0.002$, Kolmogorov–Smirnov two-tail test, $\kappa = 0.2673$). Sample traces in inset (red box: stimulation duration). Consecutive trials are overlaid. **e**, Optical stimulation in the neutral context induces significant lick suppression after training ($n = 4$ mice, 12 total trials, mean \pm s.d., $P < 0.01$, paired t -test). No lick suppression was seen before training (not significant, paired t -test). FC, fear context; NC, neutral context. **f**, Fraction of fear- and neutral-context HC and non-HC neurons recruited during before- ($n = 3$ mice, 10 total trial) and after-training optical stimulation ($n = 4$ mice, 12 total trials, $P < 0.001$, one-way ANOVA); mean, quartile, minimum and maximum are shown. $^{**}P < 0.01$, $^{***}P < 0.001$.

opsins, however, were not suitable because, as we and others have found, C1V1 (ref. 33) does not traffic well to the terminals of long-range projections, and ReaChR³⁴ (despite adequate trafficking) shows lower photocurrents and reduced spike fidelity in mammals. By extensively modifying ReaChR to include ChETA-based³⁵ and other mutations, we generated a red-shifted opsin termed bReaChES with strong photocurrents, high spike fidelity (Extended Data Fig. 10) and robust trafficking in long-range projections (Fig. 4a).

Mice were injected with GCaMP6m in CA3 and bReaChES in the AC, and implanted with a cannula above CA2/CA3 (Fig. 4a) for simultaneous one-photon stimulation of projection terminals and two-photon imaging of CA2/CA3 pyramidal neurons through the same window (Fig. 4b). To test the causal effect of the projection, we performed multiple optical-stimulation trials both before and after fear conditioning, while tracking the same neurons across contexts and days (Fig. 4c). While trial-to-trial variability existed in the number and identity of neurons activated, fear conditioning was found consistently to increase the fraction of cells that were time locked to the onset of optogenetic stimulus of the top-down projection (Fig. 4d; $P = 0.002$, Kolmogorov–Smirnov two-tail test). Finally, we directly tested recruitment of the memory-associated HC neurons by this projection. To do this, we first established that, consistent with earlier results, head-fixed mice were able to learn the contextual fear conditioning task and that stimulation of the AC–CA projection induced fear memory retrieval only after training and not before (Fig. 4e). Quantification over many trials indicated that stimulation of the AC–CA projection recruited relatively few (~5%) HC neurons in either the fear or neutral context before training, whereas there was a marked increase (~20%) in the fraction of HC neurons recruited after training; recruitment of non-HC neurons in the fear context, and recruitment of any neurons in the neutral context, remained unchanged and low (Fig. 4f). These results further demonstrated swift reorganization of the functional impact of the AC–CA projection, with preferential recruitment of HC neurons associated with the recently formed contextual fear memory. Together, these findings reveal a means by which top-down circuit influences could organize and engage with salient memory representations to enable efficient retrieval.

Discussion

We have identified a direct projection from the AC to the hippocampus with properties that mediate retrieval of recently encoded memory traces. The ability of the AC to select appropriate targets among hippocampal pyramidal neurons during context retrieval points to potential reciprocity between these two regions; the hippocampus is well positioned to inform the AC regarding context through bottom-up pathways, after which the AC can access and mobilize engrams in the hippocampus for top-down control of retrieval. The AC may also form an independent representation during training, as suggested by the speed (within 1 day post-training) with which the AC–CA projection can mediate top-down retrieval. This rapid retrieval also demonstrates that the AC is engaged early during the memory encoding process, as has been suggested^{9–11}, which, together with the previous finding of hippocampal engagement even at later time points³⁶, suggests a shift towards viewing the AC and hippocampus, and their bidirectional communication, as having shared involvement in both early and late stages of memory. The speed and specificity of the top-down retrieval also underscores likely plasticity at the AC–CA terminals for dynamic access to recently allocated neurons in the hippocampus.

We have found that this top-down control over contextual memory involves preferential recruitment of a sparse set of hippocampal neurons that are highly correlated in the local network and that tend to lead population-wide synchronous events, but only in the memory context and in a manner only seen after training. Sparsification of memory-associated networks during learning has been observed^{30,37–40}

and computationally predicted^{20,41,42}, since hierarchical networks with few extensively connected hub-like nodes, as observed here, are well suited for memory stability (random degradation of the network will probably affect non-hub nodes that are less consequential to the memory representation). But beyond memory stability, the emergence of contextual memory-specific HC neurons could facilitate efficient access to engrams, in a process suitable for memories that are strongly encoded via repetition, reward, or emotional saliency. This circuit property could be broadly relevant, helping to explain how even single cortical neurons can in some cases drive network activity and behaviour^{43–46}.

In the future, improvements in two-photon stimulation of individually targeted neurons over large three-dimensional volumes, together with simultaneous recording of network activity^{47–49}, may enable further delineation of the causal role of HC neurons during memory retrieval. Also of importance is the exploration of the molecular and structural properties of the observed HC neurons, and their inputs and outputs, through high-content methods such as single-cell RNA-sequencing and CLARITY (a method for creating labelled hybrids or composites of biomolecules in tissue covalently linked to acrylamide-related polymer hydrogels, which allows removal of unlinked tissue elements to create transparency and high-resolution accessibility to the macromolecular labels)⁵⁰. More broadly, further study of the AC–CA and other top-down projections with the circuit tracing and control methods described here may help to elucidate brain-wide regulation of memory in normal behaviour and in maladaptive states involving aberrant communication between PFC and hippocampus.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 27 May; accepted 10 August 2015.

Published online 5 October 2015.

- Kandel, E. R., Dudai, Y. & Mayford, M. R. The molecular and systems biology of memory. *Cell* **157**, 163–186 (2014).
- Han, J. H. *et al.* Neuronal competition and selection during memory formation. *Science* **316**, 457–460 (2007).
- Han, J. H. *et al.* Selective erasure of a fear memory. *Science* **323**, 1492–1496 (2009).
- Yiu, A. P. *et al.* Neurons are recruited to a memory trace based on relative neuronal excitability immediately before training. *Neuron* **83**, 722–735 (2014).
- Reijmers, L. G., Perkins, B. L., Matsuo, N. & Mayford, M. Localization of a stable neural correlate of associative memory. *Science* **317**, 1230–1233 (2007).
- Liu, X. *et al.* Optogenetic stimulation of a hippocampal engram activates fear memory recall. *Nature* **484**, 381–385 (2012).
- Garner, A. R. *et al.* Generation of a synthetic memory trace. *Science* **335**, 1513–1516 (2012).
- Ramirez, S. *et al.* Creating a false memory in the hippocampus. *Science* **341**, 387–391 (2013).
- Tse, D. *et al.* Schema-dependent gene activation and memory encoding in neocortex. *Science* **333**, 891–895 (2011).
- Lesburguères, E. *et al.* Early tagging of cortical networks is required for the formation of enduring associative memory. *Science* **331**, 924–928 (2011).
- Bero, A. W. *et al.* Early remodeling of the neocortex upon episodic memory encoding. *Proc. Natl Acad. Sci. USA* **111**, 11852–11857 (2014).
- Frankland, P. W. & Bontempi, B. The organization of recent and remote memories. *Nature Rev. Neurosci.* **6**, 119–130 (2005).
- Ressler, K. J. & Mayberg, H. S. Targeting abnormal neural circuits in mood and anxiety disorders: from the laboratory to the clinic. *Nature Neurosci.* **10**, 1116–1124 (2007).
- Taylor, S. F. *et al.* Meta-analysis of functional neuroimaging studies of emotion perception and experience in schizophrenia. *Biol. Psychiatry* **71**, 136–145 (2012).
- Wilson, S. J., Sayette, M. A. & Fiez, J. A. Prefrontal responses to drug cues: a neurocognitive analysis. *Nature Neurosci.* **7**, 211–214 (2004).
- Nadel, L. & Moscovitch, M. Memory consolidation, retrograde amnesia and the hippocampal complex. *Curr. Opin. Neurobiol.* **7**, 217–227 (1997).
- Winocur, G., Moscovitch, M. & Bontempi, B. Memory formation and long-term retention in humans and animals: convergence towards a transformation account of hippocampal-neocortical interactions. *Neuropsychologia* **48**, 2339–2356 (2010).
- Strogatz, S. H. Exploring complex networks. *Nature* **410**, 268–276 (2001).
- Barabási, A. L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nature Rev. Genet.* **5**, 101–113 (2004).

20. Bullmore, E. & Sporns, O. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Rev. Neurosci.* **10**, 186–198 (2009).
21. Hermundstad, A. M., Brown, K. S., Bassett, D. S. & Carlson, J. M. Learning, memory, and the role of neural network architecture. *PLOS Comput. Biol.* **7**, e1002063 (2011).
22. Wickersham, I. R., Finke, S., Conzelmann, K. K. & Callaway, E. M. Retrograde neuronal tracing with a deletion-mutant rabies virus. *Nature Methods* **4**, 47–49 (2007).
23. Soudais, C., Laplace-Builhe, C., Kiss, K. & Kremer, E. J. Preferential transduction of neurons by canine adenovirus vectors and their efficient retrograde transport *in vivo*. *FASEB J.* **15**, 2283–2285 (2001).
24. Dombeck, D. A., Harvey, C. D., Tian, L., Looger, L. L. & Tank, D. W. Functional imaging of hippocampal place cells at cellular resolution during virtual navigation. *Nature Neurosci.* **13**, 1433–1440 (2010).
25. Mahoney, W. J. & Ayres, J. J. B. One-trial simultaneous and backward fear conditioning as reflected in conditioned suppression of licking in rats. *Anim. Learn. Behav.* **4**, 357–362 (1976).
26. Bouton, M. E. & Bolles, R. C. Conditioned fear assessed by freezing and by the suppression of three different baselines. *Anim. Learn. Behav.* **8**, 429–434 (1980).
27. Lovett-Barron, M. *et al.* Dendritic inhibition in the hippocampus supports fear learning. *Science* **343**, 857–863 (2014).
28. Chen, T. W. *et al.* Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **499**, 295–300 (2013).
29. Cheng, S. & Frank, L. M. New experiences enhance coordinated neural activity in the hippocampus. *Neuron* **57**, 303–313 (2008).
30. Komiyama, T. *et al.* Learning-related fine-scale specificity imaged in motor cortex circuits of behaving mice. *Nature* **464**, 1182–1186 (2010).
31. Modi, M. N., Dhawale, A. K. & Bhalla, U. S. CA1 cell activity sequences emerge after reorganization of network correlation structure during associative learning. *eLife* **3**, e01982 (2014).
32. Vogelstein, J. T. *et al.* Fast nonnegative deconvolution for spike train inference from population calcium imaging. *J. Neurophysiol.* **104**, 3691–3704 (2010).
33. Yizhar, O. *et al.* Neocortical excitation/inhibition balance in information processing and social dysfunction. *Nature* **477**, 171–178 (2011).
34. Lin, J. Y., Knutsen, P. M., Muller, A., Kleinfeld, D. & Tsien, R. Y. ReaChR: a red-shifted variant of channelrhodopsin enables deep transcranial optogenetic excitation. *Nature Neurosci.* **16**, 1499–1508 (2013).
35. Gunaydin, L. A. *et al.* Ultrafast optogenetic control. *Nature Neurosci.* **13**, 387–392 (2010).
36. Goshen, I. *et al.* Dynamics of retrieval strategies for remote memories. *Cell* **147**, 678–689 (2011).
37. Vinje, W. E. & Gallant, J. L. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science* **287**, 1273–1276 (2000).
38. Olshausen, B. A. & Field, D. J. Sparse coding of sensory inputs. *Curr. Opin. Neurobiol.* **14**, 481–487 (2004).
39. Yassin, L. *et al.* An embedded subnetwork of highly active neurons in the neocortex. *Neuron* **68**, 1043–1050 (2010).
40. Gdalyahu, A. *et al.* Associative fear learning enhances sparse network coding in primary sensory cortex. *Neuron* **75**, 121–132 (2012).
41. Buzsáki, G., Geisler, C., Henze, D. A. & Wang, X.-J. Interneuron Diversity series: circuit complexity and axon wiring economy of cortical interneurons. *Trends Neurosci.* **27**, 186–193 (2004).
42. Perin, R., Berger, T. K. & Markram, H. A synaptic organizing principle for cortical neuronal groups. *Proc. Natl Acad. Sci. USA* **108**, 5419–5424 (2011).
43. Bonifazi, P. *et al.* GABAergic hub neurons orchestrate synchrony in developing hippocampal networks. *Science* **326**, 1419–1424 (2009).
44. Brecht, M., Schneider, M., Sakmann, B. & Margrie, T. W. Whisker movements evoked by stimulation of single pyramidal cells in rat motor cortex. *Nature* **427**, 704–710 (2004).
45. Houweling, A. R. & Brecht, M. Behavioural report of single neuron stimulation in somatosensory cortex. *Nature* **451**, 65–68 (2008).
46. Li, C. Y., Poo, M.-M. & Dan, Y. Burst spiking of a single cortical neuron modifies global brain state. *Science* **324**, 643–646 (2009).
47. Prakash, R. *et al.* Two-photon optogenetic toolbox for fast inhibition, excitation and bistable modulation. *Nature Methods* **9**, 1171–1179 (2012).
48. Rickgauer, J. P., Deisseroth, K. & Tank, D. W. Simultaneous cellular-resolution optical perturbation and imaging of place cell firing fields. *Nature Neurosci.* **17**, 1816–1824 (2014).
49. Packer, A. M., Russell, L. E., Dagleish, H. W. & Häusser, M. Simultaneous all-optical manipulation and recording of neural circuit activity with cellular resolution *in vivo*. *Nature Methods* **12**, 140–146 (2015).
50. Chung, K. *et al.* Structural and molecular interrogation of intact biological systems. *Nature* **497**, 332–337 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank B. K. Lim for the rabies virus and S. Pak for assistance with histology. We thank the entire Deisseroth laboratory for thoughtful comments, with particular gratitude to M. Lovett-Baron, A. Andalman and W. Allen for their helpful discussions throughout. K.D. is supported by the Defense Advanced Research Projects Agency Neuro-FAST program, National Institute of Mental Health, National Institute on Drug Abuse, National Science Foundation, the Simons Foundation, the Tarlton Foundation, the Wieggers Family Fund, the Nancy and James Grosfeld Foundation, the H.L. Snyder Medical Foundation, and the Samuel and Betsy Reeves Fund. This work is supported by an Ellison Life Sciences Research Foundation (LSRF) fellowship (P.R.), a Simons LSRF fellowship (J.H.M.), the German Academic Exchange Service DAAD (A.B.) and the Fidelity Foundation (S.Y.L.). All tools and methods are distributed and supported freely (<http://www.optogenetics.org>).

Author Contributions P.R. and K.D. designed the experiments. P.R. performed anatomical tracing, optogenetic behaviour, virtual reality behaviour, hippocampal cranial window surgeries and *in vivo* imaging experiments, and collected all associated data. S.S. wrote custom code to extract neural sources and performed computational analysis on all of the *in vivo* calcium imaging datasets. P.R. and J.H.M. collected data from simultaneous one-photon stimulation and two-photon imaging experiments, and J.H.M. and S.S. analysed those data. E.F. and S.Y.L. performed patch electrophysiology experiments. C.K.K. designed the virtual reality infrastructure. S.Y.L., A.B. and C.R. designed and tested the bReaChES opsin. A.J. performed injections and fibre implant surgeries. M.L. assisted with behaviour. C.L. assisted with cranial window surgeries and statistical analysis. P.R. and K.D. wrote the paper; K.D. supervised all aspects of the work. All authors discussed findings, edited and contributed to the final version of the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to K.D. (deissero@stanford.edu).

METHODS

Animals. Wild-type C57Bl6/J male mice were group housed three to five to a cage and kept on a reverse 12 h light/dark cycle with *ad libitum* food and water (except in virtual reality behaviour experiments, for which water was restricted; details later). Experimental protocols were approved by Stanford University Institutional Animal Care and Use Committee (IACUC) and meet the guidelines of the National Institutes of Health guide for the Care and Use of Laboratory Animals. The target number of subjects used in each experiment was determined based on numbers reported in published studies. No statistical methods were used to predetermine sample size.

Anatomical tracing and histology. Viral injections were carried out under protocols approved by Stanford University IACUC and were performed in mice anaesthetized with 1–2% isoflurane using a stereotaxic apparatus (Kopf Instruments). For retrograde tracing, 4–5-week-old wild-type male mice were injected slowly (50 nl min^{-1}) with small amounts (200 nl) of highly concentrated glycoprotein-deleted rabies virus tagged with tdTomato (RV-tdTomato)²² in the dorsal hippocampus (A/P: -1.5 mm ; M/L: $+1.75 \text{ mm}$; D/V: -1.8 mm) with a $1 \mu\text{l}$ Hamilton syringe and a 35-gauge bevelled needle (World Precision Instruments) under the control of a UMP3 syringe pump (WPI). Following injections, the incisions were closed using Vetbond tissue adhesive (Fischer), and mice were allowed to recover and were housed for 5 days to allow for expression before their brains were collected for histological analysis. In the case of anterograde tracing, 4–5-week-old wild-type male mice were injected (150 nl min^{-1}) with 500 nl of AAV5-CaMKII α ::eYFP (titre: $2 \times 10^{12} \text{ vg ml}^{-1}$) in dorsal anterior cingulate (A/P: $+1$; M/L: -0.35 ; D/V: $+1.2$) and were housed for 30 days to allow for expression in terminals before collection of brains for histological analysis.

For histological analysis, injected mice were transcardially perfused with ice-cold $1 \times \text{PBS}$, immediately followed by perfusion of 4% paraformaldehyde (PFA). Brains were fixed overnight in PFA, then transferred to a 30% sucrose/PBS solution. Coronal sections of either $40 \mu\text{m}$ (for retrograde tracing with RV) prepared using a freezing microtome (Leica) or $300 \mu\text{m}$ (for anterograde tracing with AAV5) prepared using a vibratome (Leica) were collected and stored in a cryoprotectant solution (25% glycerol, 30% ethylene glycol, in PBS) until further processing. For DAPI staining, slices were washed in PBS, incubated for 20 min with DAPI at $1:50,000$, washed again in PBS, then mounted with PVA-DABCO (Sigma). A scanning confocal microscope (TCS SP5, Leica) and LAS AF software (Leica) was used to obtain and analyse images.

Acute slice electrophysiology of AC-CA synapses. Acute brain slices were prepared from mice 6–8 weeks following viral injection with AAV5-CaMKII α ::Chr2(H134R)-eYFP, to allow sufficient time for channelrhodopsin to express in axon terminals. After lethal anaesthesia, mice were transcardially perfused with cold sucrose slicing solution (see later) before decapitation, following which the brain was rapidly extracted and submerged in ice-cold sucrose-based slicing solution (234 mM sucrose, 26 mM NaHCO_3 , 11 mM glucose, 10 mM $\text{MgSO}_4 \cdot 7\text{H}_2\text{O}$, 2.5 KCl, 1.25 mM $\text{NaH}_2\text{PO}_4 \cdot \text{H}_2\text{O}$, 0.5 mM $\text{CaCl}_2 \cdot 2\text{H}_2\text{O}$). Coronal hippocampal slices ($300 \mu\text{m}$ thick) were cut on a Leica vibratome (Leica VT1000S) in sucrose solution and then submerged in a hypertonic recovery solution (artificial cerebrospinal fluid (ACSF) at an 8% increased osmolality) at 33°C for 15 min before being transferred to standard ACSF (123 mM NaCl, 26 mM NaHCO_3 , 11 mM glucose, 3 mM KCl, 2 mM $\text{CaCl}_2 \cdot 2\text{H}_2\text{O}$, 1.25 mM $\text{NaH}_2\text{PO}_4 \cdot \text{H}_2\text{O}$, 1 mM $\text{MgCl}_2 \cdot 6\text{H}_2\text{O}$) for a further 45 min at 33°C , at which point they were transferred to room temperature.

Whole-cell patch-clamp recordings from CA3/CA1 hippocampal neurons were performed on an upright Leica DM-LFSA microscope. Borosilicate glass (Sutter Instruments) pipette resistances were pulled to 3–6 M Ω and filled with potassium gluconate intracellular solution (130 mM Kgluconate, 10 mM KCl, 10 mM HEPES, 10 mM EGTA, 2 mM MgCl_2 , pH adjusted with KOH to 7.3). Voltage and current-clamp recordings were performed using pClamp (Axon Instruments). Cells with leak current greater than -200 pA or series resistance greater than 35 M Ω were excluded. Light stimulation was performed using a 300 W DG-4 lamp (Sutter Instruments) with an external filter for blue light (wavelength in nm/bandwidth in nm: 470/20). Light pulses (2–5 ms pulse width) were delivered through a $\times 40$, 0.8 NA water-immersion objective at $4\text{--}10 \text{ mW mm}^{-2}$ light power density. Latencies were measured as light pulse start to EPSC initiation.

Optogenetics and behaviour. After injection with the indicated virus (for example, CAV or RV, expressing Chr2, eNpHR3.0, or eYFP) at the appropriate location (for example, cingulate, hippocampus, or medial septum), as described in Fig. 2 and Extended Data Figs 2 and 4, 5-week-old wild-type male mice were implanted with implantable fibre-optic lightguides (IFLs) consisting of a 2.5-mm-diameter metal ferrule with 0.22 NA and a 200- μm -thick protruding cleaved bare optic fibre cut to the desired length (Thorlabs) as previously described³⁶, either at the injection site (typically $\sim 0.2 \text{ mm}$ dorsal to the injection site) or at the term-

inals for stimulation experiments as indicated in the figure legends. For inhibition experiments, dual fibre-optic cannulas of 200 μm thickness and 0.22 NA spaced 0.7 mm apart were used to target anterior cingulate bilaterally, and two-ferrule cannulas spaced 3 mm apart were used to target hippocampus bilaterally. Mice were typically allowed to recover and housed for 1 month to allow for adequate expression before behavioural testing. All animals undergoing behavioural experiments were acclimated to a 12 h reverse light/dark cycle, handled for several days, and before behavioural testing, were acclimated to the room in which experiments were to be conducted for at least 30 min.

The fear conditioning apparatus consisted of a square conditioning cage ($18 \times 18 \times 30 \text{ cm}$) with a grid floor wired to a shock generator and a scrambler, surrounded by an acoustic chamber (Coulbourn Instruments). The apparatus was modified to enable light delivery during retrieval testing. Contextual fear conditioning was performed by placing mice in the conditioning cage (visual cues: bare walls; tactile cues: grid floor; odour cues: 70% ethanol) for 6 min, while receiving four 2 s shock pulses of 7 mA each at 1 min intervals, with the first shock presented 2 min after placing the mouse in the conditioning context. A fraction of animals of the same cohort were not fear conditioned, and instead served as a control group that were just exposed to the conditioning context for the same amount of time (6 min) but did not receive any associated shocks. The following day, all mice were tested in a different 'neutral' cage (visual cues: coloured shapes; tactile cues: smooth paper towel covered plexiglass floor; odour cues: 1% acetic acid) for light-mediated fear retrieval.

For stimulation experiments, optical stimulation through the fibre-optic connector was administered by delivering light through a patch-cord connected to a 473 nm laser in 30 s light-on/1 min light-off sessions. During light-on sessions, stimulation was delivered at 20 Hz, 15 ms pulses, with 8–10 mW power at the fibre tip. On the third day, all mice were then returned to the original conditioning context for 2.5 min to assess intact natural fear memory retrieval. In some cases, subsequent extinction of fear memory was performed by placing mice in the original conditioning chamber for three consecutive days, for 5 min each, without shock. Light-induced fear retrieval was then tested in the neutral context 24 h following the last extinction training session. Subsequent reinstatement was performed by again placing the animals back in the conditioning context for one 6 min interval and providing four 2 s shock pulses of 7 mA each at 1 min intervals. A final light-induced fear retrieval testing was performed 24 h later as described earlier.

For loss of function experiments, optical inhibition through a fibre-optic connector was administered by delivering light through a dual patch-cord connected to a 589 nm laser. Constant light at 8–10 mW was used at the fibre tip to deliver inhibition either at cell bodies or terminals. On the first day, both eNpHR3.0 and eYFP control groups were trained to contextual fear conditioning as described earlier, and on the second day, mice were allowed to perform retrieval as usual during light off for the first 2 min, to assess baseline freezing in each animal. Then light was turned on for the next 30 s (not longer, as the potential for extinction related unfreezing could confound light-related unfreezing at time points succeeding the typical 2–3-min retrieval protocol). Freezing scores during the 30 s light sessions were compared with the per cent freezing during 30 s of the immediately preceding light-off sessions. On the third day, all mice underwent retrieval in the conditioning context for 2 min with light off to test for reversal of light-induced behaviour. After context conditioning and retrieval, all mice subsequently underwent auditory-cued conditioning (cued conditioning was done separately from context conditioning to ensure robust conditioning to both context and cue, since when performed together, mice often develop robust conditioning to tone (the more salient cue) and only weak conditioning to context). To perform auditory-cued fear conditioning, mice were placed in a different context (with coloured shapes as visual cues and a smooth floor), for 6 min, where after the first 2 min, four 20 s auditory cues consisting of 2.9 kHz tone was played at 1 min intervals, each followed by a 2 s 7 mA shock. Retrieval on the subsequent day was performed by presenting the tone four times (two during light off and two during light on) at 1 min intervals and per cent freezing was assessed during the 20 s post-tone compared with the immediately preceding 20 s during tone, for both light-off and light-on conditions. Latency measures were performed as separate experiments, using the same cohorts; after finishing contextual and cued conditioning, these mice were retrained (contextually fear conditioned) to the first conditioning context. On the following day, 2 min retrieval was performed in the conditioning context with light on the entire time to test for latency to freezing, where latency was defined as the first instance in time that the animal was immobile for 5 consecutive seconds. Freezing in all experiments was scored by an experimenter blinded to the treatment group. Randomization of animals to experimental and control groups was performed by an experimenter with no explicit randomization algorithm used. All of the results were analysed by Student's *t*-test or two-way ANOVA, followed by post-hoc tests, as applicable.

Hippocampal cranial window. C57BL/6J male mice were injected with 500 nl of AAVdj-CaMKII α ::GCaMP6m in CA3 (A/P: -1.7, M/L: +1.9, D/V: -1.9) and allowed to recover for at least 1 week before surgical implantation of a cranial window above CA2/CA3 for optical access similar to previously described hippocampal preps²⁴. Briefly, mice were injected with 80 mg kg⁻¹/6 mg kg⁻¹ of ketamine/xylazine intraperitoneally, and maintained under 1.0–2.0% isoflurane throughout. For optimal window placement to access CA2/CA3, the mouse's head was angled during surgery such that the skull location at the CA2/CA3 injection site was level and exactly perpendicular to dorsal views of the head. A circular titanium headplate (7 mm in diameter) was centred over CA2/CA3 and adhered to the skull with adhesive cement (Metabond; Parkell) and a ~3 mm craniotomy was made in the centre using a trephine (Fisher). Parts of cortical region S1 and of parietal association cortex were vacuum-aspirated, with care taken to avoid the ventricle, until white matter was visible above the hippocampus. Vacuum aspiration was done with a 27-gauge blunt needle while irrigating with chilled 1× PBS. The top layer of white matter above the hippocampus was further removed by vacuum aspiration with a 31-gauge blunt needle, but care was taken to preserve deep layers of external capsule and the alveus (to preserve afferents and efferents to hippocampus). A forceps was used to manually insert a cylindrical borosilicate glass implant until the floor of the implant rested against the hippocampus. The implant was constructed from a 3.0-mm-diameter glass capillary tube (Friedrich & Dimmock) custom cut to 1.5 mm length, adhered on one end to a 3.0 mm diameter coverslip of #0 thickness (Warner Instruments) using UV-curing optical glue (Norland Products). The top of the implant extruding from the craniotomy was then secured to the skull using Metabond adhesive cement. After surgery, mice were given 5 mg kg⁻¹ carprofen subcutaneously and allowed to recover for at least 1 week before behaviour training.

To ensure that the above manipulations (including GCaMP6m virus injection into CA3, GCaMP6m expression, and surgical excavation of certain regions of cortex) did not affect normal physiological properties of the hippocampus, we performed control experiments to assess Ca²⁺-dependent physiology in weakly versus strongly expressing CA3 neurons *in vitro*, spontaneous activity in weakly versus strongly expressing CA3 neurons *in vivo*, and behavioural measurements before and after placement of the cannula (Extended Data Fig. 3).

Virtual reality behaviour. We used a custom built virtual reality environment, modified from previously reported versions^{24,51}. A 200-mm-diameter styrofoam ball (Graham Sweet Studios) was axially fixed with a 6-mm-diameter assembly rod (Thorlabs) passing through the centre of the ball and resting on 90° post holders (Thorlabs) at each end, allowing free forward and backward rotation of the ball. Mice were head-fixed in place above the centre of the ball using a head-plate mount⁵². Virtual environments were designed in game development software Unity3d (<http://www.unity3d.com>). The virtual environment was displayed by back-projection onto projector screen fabric stretched over a clear acrylic hemisphere with a 14-inch diameter placed ~20 cm in front of the centre of the mouse. The screen encompasses ~220° of the mouse's field of view. The virtual environment was back-projected onto this screen using two laser-scanning projectors (Microvision), each projector covering one half of the screen. To create a flat image on the three-dimensional screen, we warped the two-dimensional image of the virtual environment using video manipulation software (Madmapper). The game engine allowed scripts written in JavaScript or C# to trigger external events based on the mouse's interactions with the virtual environment by communicating over a TCP socket to custom Python control software. A LabJackU6 (<http://labjack.com>) was used to time-lock virtual environment events and imaging frame times, to record mouse licking behaviour with incoming TTL pulses from the lickometer (Island Motion), and to send TTL pulses to deliver solenoid-gated water rewards (delivered from a gravity-assisted syringe attached to tubing connected to the lickometer) and aversive air puffs (from a compressed air tank to a tube ending in a pipette tip facing the mouse's snout). Tactile and odour cues were fixed directly to each of two Styrofoam balls representing the two separate contexts. Auditory stimuli were presented through speakers situated behind the animal. The mouse's movements on the ball were recorded using an optical computer mouse (Logitech) that interfaced with the virtual environment software.

For fear conditioning in the virtual environment, mice were water restricted (>80% pre-deprivation weight) and habituated to handling, head-fixation, and the virtual environment for at least 2 weeks, with free access to small water rewards (~0.5 µl per 10 licks) while on the ball. By the end of 2 weeks (one 5-min session per day), mice appeared comfortable and alert on the ball. After habituation, mice underwent a 4-day fear conditioning training and testing protocol. On day 1, mice were exposed to two contexts that differed in visual (blue triangles versus pink vertical stripes), tactile (smooth side of Velcro versus sharp side of Velcro fixed onto running ball), odourant (acetic acid versus ethanol), and auditory cues (8 kHz phasic tone versus 3 kHz pure tone) for 5 min each. On day 2, mice were provided with 8 aversive air puffs to the snout (500 ms, 10 psi) at

randomly timed intervals throughout the 5 min while in the fear context, but not while in the neutral context for 5 min. On days 3 and 30, mice were placed back in each of the two contexts for 5 min for retrieval.

Imaging. Five mice were imaged on all days, in 5-min sessions, during exposure, training, and retrieval. We used a resonant galvanometer two-photon microscope (Prairie Technologies). We used the genetically encoded calcium indicator GCaMP6m in all experiments (GCaMP6m was amplified from Addgene plasmid #40754 by PCR and subcloned into an AAV backbone under the control of the CaMKII α promoter.) All experiments were performed using a Coherent Ultra II Ti:Sapphire pulsed laser tuned to 920 nm to excite GCaMP6m through a ×20 0.5 LUMPlanFL/N (Olympus) water-immersion objective interfacing with the implanted cannula through a few drops of distilled water. Fluorescence was detected through gallium arsenide phosphide (GaAsP) photomultiplier tubes (PMTs) using the PrairieView acquisition software. High speed z stacks were collected in the green channel (using a 520/44 bandpass filter, Semrock) at 512 × 512 pixels covering each x-y plane of 500 µm × 500 µm over a depth of ~100 µm (3–7 z slices ~10–20 µm apart) by coupling the 30 Hz rapid resonant scanning (x-y) to a Z-piezo to achieve ~6 Hz per volume.

Data analysis. Later, we describe the methods to extract cells (pre-processing), obtain cellular-level activity ($\Delta F/F$) measures (processing), and evaluate population-level activity measures (post-processing). In statistical analysis of the post-processed data, both parametric and non-parametric tests were employed as appropriate. In cases where normality could not be assessed (low sample sizes), we ensured that there were no significant outliers (by Grubbs' test) and that the variance between groups was not significantly different (by Levene's Test).

Pre-processing (cell extraction). Time series data sets were x-y motion corrected with ImageJ plug-in Stack Reg using rigid body transformations. Cell extraction was then performed sequentially, by first computing cell segments automatically followed by manual quality control for missed cells, non-cells, or conjoined cells. For initial automatic extraction, we used a metric based on image threshold intensity, variance and skewness. Images with high contrast-to-noise ratio, wherein clear thresholds in maximum intensity separated cells and background, were fully segmented with the former. In the remainder of cases, cells were distinguished from background based on standard deviation across time (high for active cells), or skewness (asymmetry) in intensity across time⁵³. This resulted in a general mathematical criterion to define cell-masks at each voxel location (i, j, k):

$$M(i, j, k) = \alpha_F I(F_{\max}(i, j, k) > F_c) + \beta_F I(\sigma_F(i, j, k) > \sigma_c) + \gamma_F I(s_F(i, j, k) > s_c)$$

where I is the indicator function (= 1 if the condition is satisfied); $\sigma_F(i, j, k)$ is the standard deviation of intensity over time defined as

$$\sigma_F^2(i, j, k) = E[(F(i, j, k, t) - \bar{F})^2]$$

and skewness is defined as

$$s_F(i, j, k) = E\left[\left(\frac{F(i, j, k, t) - \bar{F}}{\sigma_F(i, j, k)}\right)^3\right]$$

E is the expectation operator; F_c , σ_c and s_c represent cut-offs for image intensity, standard deviation and skewness respectively. Coefficients α_F , β_F and γ_F are chosen on an image-specific basis; if thresholding is sufficient β_F and γ_F are chosen to be zero, otherwise coefficients are iterated to obtain a cell mask containing the largest population of active cells (evaluated by inspection).

Automatic cell extraction was then followed by manual cell-by-cell curation to identify cells that were not extracted using the automated algorithm. This occurs when cell boundaries may not be captured due to non-translational motion artefact in the original imaging, and/or lack of clear cut-offs F_c , σ_c and s_c differentiating cell and background. For these cases, cell detection is performed with a manual editing step involving comparison of the automated cell-mask to the raw image data, and by using a Gaussian filter was applied on the edited image to smooth edges, and edge-detection⁵⁴ was used to define cell boundaries. The interior of the resulting cells were filled, and the final cell masks were eroded to minimize contamination from neuropil signal. Each cell was labelled with a unique cell identifier for the next stage; custom-written MATLAB scripts were used for all steps, and are available on request.

Processing. Calculation of $\Delta F/F$. For each cell identified in step 1, the intensity value F was obtained by averaging over all pixels inside the ROI to compute a space-averaged value \bar{F} for each frame (corresponding to a single time point). These are used to define $\Delta F/F$ in each cell as

$$\frac{\Delta F}{F} = \frac{F - \bar{F}_{\text{baseline}}}{\bar{F}_{\text{baseline}}}$$

where $\bar{F}_{\text{baseline}}$ is the baseline fluorescence, calculated as the mean of the fluorescence values for a given cell, continuously acquired over a 20 s moving time window to account for slow time-scale changes in fluorescence. Given the sparse firing of neurons in our data set, the mean served as an accurate estimate of baseline activity (fluorescence). Furthermore, the main results of the study were not influenced by using the median or 8th percentile as the baseline (and correlations were independent of baseline definition).

Statistical analysis of neuronal responses. We used an approach similar to that outlined previously²⁴ to identify significant transients in each neuron, as well as to estimate and remove effects that may be related to motion artefacts. Briefly, to estimate the occurrence rate of potential motion-related fluorescence changes in the signal, all negative deflections in the $\Delta F/F$ trace were assumed to be due to motion. Because motion-related fluorescence changes should be equally likely to generate positive- or negative-going changes, positive and negative deflections in the $\Delta F/F$ curve that are attributable to motion should occur at the same frequency and can be subtracted out of the signal by using the rate of occurrence of the negative-going transients as an estimate of the rate of motion-related positive-going transients.

To determine statistically significant transients, we first calculated an estimate of the noise for each cell using an iterative approach: (1) initialize a cut-off value that separates signal and noise, (2) calculate the standard deviation (σ) of all $\Delta F/F$ values that fall below the cut-off, and (3) compare 3σ to the cut-off. In this analysis, the goal is to find an estimate of standard deviation (σ) of the noise, defined for time periods that are unlikely to contain neural events (that is, using the iterative approach to estimate the σ of the noise, rather than calculate standard deviation for the entire time epoch, which would contain real events). For each iteration of the analysis, if $|\text{cut-off} - 3\sigma| < \text{tolerance}$, the program terminates (where tolerance = 0.02). If $\text{cut-off} > 3\sigma$, the program increases the cut-off by 10% and goes back to step 1. If $\text{cut-off} < 3\sigma$, it reduces the cut-off by 10% and goes back to step 1. This approach helped ensure that neuronal activity-generated events in $\Delta F/F$ are not included in the estimation of noise and avoided the need for manually selecting epoch intervals on a cell-by-cell basis that did not contain an event in order to estimate noise.

Subsequently, we analysed positive- and negative-going transients to further determine the false positive rate. Transient onsets are defined as the times when the $\Delta F/F$ exceeds 2σ and offset is defined as the time at when a given transient falls below 0.5σ . A histogram of the number of transients that exist for each σ threshold value (that is, $>2\sigma$, $>3\sigma$, $>4\sigma$), for various durations, is extracted, where negative-going transients are to the left of the ordinate and plotted in red (Extended Data Fig. 4c–f). The ratio of the number of negative to positive going transients is calculated for different transient durations across three amplitude levels (2σ , 3σ , 4σ), and serves as our estimate of false positive rate. Following from the reasoning described earlier, this ratio will be 50% when the motion-based noise significantly exceeds the signal. We plot the false positive ratio for the different scenarios described earlier, and choose the amplitude (in σ) and duration cut off (Extended Data Fig. 4c–f) needed to reduce the false positive rate to below 5%. As mentioned previously³, it is important to note that this estimate of noise represents an upper bound, and could be influenced by other sources of noise apart from motion (that is, photon shot noise).

Calculation of correlation coefficients between neuron pairs. The Pearson correlation coefficient was calculated between each pair of cells, c_a and c_b , as

$$\rho_{c_a, c_b} = \frac{E((c_a - \bar{c}_a)(c_b - \bar{c}_b))}{\sigma_{c_a} \sigma_{c_b}}$$

This metric measures linear dependence between signals in the two cells, and is invariant with respect to scaling or amplitude translation of the cell signals. We define a matrix of correlation coefficients of size $N_{\text{cells}} \times N_{\text{cells}}$ wherein each entry corresponds to correlation between the cells identified by the corresponding row and column. To avoid accumulation in correlated signal due to slow drifts (for example, the long decay curve of GCaMP6m), we set all $\Delta F/F$ values lying outside the window of a significant transient (as defined earlier) to 0.

Post-processing. Histogram of cell activity correlations. The property of high correlation (HC) was tested for in each neuron by finding the number of correlated neurons with which the Pearson's correlation coefficient was above 0.3 (a Pearson correlation cut-off of 0.3 was used as a conservative estimate of connectivity since previous studies using *in vivo* two-photon calcium imaging followed by paired whole-cell recordings reported a greater than 50% chance of connectivity when correlations of Ca^{2+} signals exceeded 0.3 *in vivo*)^{55,56}. Histograms were obtained by binning this number across neurons in steps of 5 and calculating the number of neurons that fell into each bin, with the resulting histogram representing the degree distribution of all neurons in the network.

HC neurons were defined as those neurons that had more correlated partners than that of the average neuron in the same volume by >1 standard deviation.

Optimally separating hyperplane. To identify network population activity measures that best distinguished fear and neutral contexts, we used a space of graph theoretic parameters (described later), which together can be used to define an optimally separating hyperplane between the two contexts. Mathematically, this is posed as a constrained optimization problem, with the objective function seeking to maximize the sum of distances of the hyperplane to the nearest data points in each context, and the constraint being that the hyperplane separates the two contexts. This constrained optimization problem was solved using Lagrange multipliers.

Synchrony and quantification of lead-lag. To analyse the spontaneous activity of the entire network, we computed the onset and duration of each activity transient (where event onsets and offsets are calculated as described earlier) for each neuron, and then combined transients from all cells into raster plots and collapsed these raster plots into activity histograms, which indicated the percentage of active cells as a function of time.

To identify epochs of synchronous activity that included more active cells than would be expected by chance at each frame, we used interval reshuffling (randomly reordering of intervals between events for each cell), performed 1,000 times for each mouse in each context, such that a surrogate histogram was constructed for each reshuffling. The threshold percentage of active neurons corresponding to a significance level of $P < 0.05$ (appearing only in 5% of histograms) was taken to be the per cent of coactive cells required in a single frame to be considered a synchronous event, and this threshold ranged between 2.5% and 5% active neurons per frame across all mice and fields of view. At least three consecutive frames with activity above the significance threshold were required to be considered a synchronous event, and all subsequent contiguous frames above this threshold were grouped together into the same synchronous event. To plot the cumulative distribution function of event onsets for HC and non-HC neurons during synchronous events, all synchronous events across all mice were identified, and the onset times of HC versus non-HC neurons were binned per frame and plotted cumulatively as a function of the percentage of time elapsed during the synchrony window.

To quantify whether the activity of HC neurons was leading or lagging their correlated pairs, the event onset of the HC neurons (defined as the first instance when the signal exceeded 3σ for time consecutive frames) was first fixed at $t = 0$. The event onsets of all correlated pairs were then binned into 0.167 s time windows immediately preceding or succeeding the onset of the hub neuron at $t = 0$.

PCA. PCA was used to describe and visualize population activity of all neurons over time in each context. This was done by transforming the $\frac{\Delta F}{F}$ of each cell (typically ~500 cells per mouse per context), over all time points in a given context (typically 1,800 frames) to a different coordinate system characterized by linearly independent eigenvectors, where each eigenvector represents a weighted combination of the different cells. Eigenvalues were sorted in decreasing order to reveal the most energetic (contributory) eigenvectors as well as the magnitude of their overall contribution. PCA was performed using eigenvalue decomposition of the correlation matrix. The corresponding eigenvalues and eigenvectors were calculated using custom MATLAB scripts.

Estimation of graph-theoretic parameters. An undirected graph is defined based on the cell correlations in the population. An edge, E , is defined between neurons if they are correlated beyond the threshold described earlier. The undirected neuronal graph $G(V, E)$ is defined using all the cells, which are denoted by V (vertices), and E (edges). Mean and maximum cell correlations are calculated using aggregate average and the maxima over all of the measured correlations. We also fit an exponential distribution between n_{corr} (the number of correlations) and n_{freq} (the degree distribution described earlier) to quantify how closely the graph mimics small world networks, which are characterized by a power law degree distribution

$$n_{\text{freq}} = a(n_{\text{corr}})^{-b}$$

for which the power law parameters, a and b , are calculated by transforming the above equation into a logarithmic scale and performing a minimum least-squares fit.

A neighbourhood is defined for each cell as a sphere of radius 30 μm . The clustering coefficient for a vertex is defined as the ratio of number of edges within its neighbourhood to the maximum number of connections possible. If there are k nodes in the neighbourhood, $k(k-1)/2$ is the maximum number of possible connections⁹. The clustering coefficient of the entire network is defined as the mean clustering coefficient across all vertices. The mean path length is defined as the average path between any two randomly selected vertices of the graph. The

mean path length (mpl) is calculated by first constructing an adjacency matrix, which is an $n_{\text{cell}} \times n_{\text{cell}}$ matrix, and all correlated vertex pairs are given a value of one in the corresponding row and column, and zero otherwise. The minimum path from i to j can be recursively calculated using

$$\text{mpl}_{i,j} = \min(\text{mpl}_{i,k} + \text{mpl}_{k,j})$$

Small-world networks are characterized by high clustering coefficient and low mean path length, quantified using the ratio of clustering coefficient to mean path length, where each term is normalized to a purely random graph with the same number of vertices. Betweenness centrality is a measure of the centrality of nodes in the network, and indicates how central a node is to communication between all pairs of node. Betweenness centrality is computed by calculating all possible paths between two nodes and calculating the number of those that pass through a given node. Strength of a graph quantifies how strongly different subcomponents of a graph are connected and is a measure of resistance of the graph to attack on its edges. Let $P = (V_1, V_2 \dots V_n)$ denote all possible partitions of the graph G into a mutually exclusive set of vertices $V_1, V_2 \dots V_m$ such that the union of all the vertices is V . Let E_r denote the number of edges that needs to be removed from G to create the partition P . Then the strength is defined as $s = \min \frac{E_r}{n-1}$, where the minima are calculated over all possible partitions P . In other words, the strength quantifies how to remove minimal edges to create maximal separation among vertices of the graph. The strength is calculated using MATLAB code based on algorithms described previously^{57,58}.

Fast non-negative deconvolution algorithm implementation. Deconvolution algorithms enable the estimation of spike rate trains from fluorescence data. Here, we use deconvolution to estimate activity-event onset, not to detect single spikes, since GCaMP6m is assumed to neither have the linear response kinetics nor the sensitivity needed to detect single spikes from bursting neurons in the hippocampus. We used this analysis to help confirm our main results regarding synchronous events and timing of highly correlated neurons, since these analyses offer an alternative method to identify event onsets, while helping to remove noise (for example, long Ca^{2+} signal decays) from the analyses.

Many deconvolution algorithms exist. Early methods to deconvolve fluorescence data used either thresholding to infer event onset⁵⁹ or optimizations to match a chosen spike profile⁶⁰. More robust algorithms such as the Wiener linear filter are promising⁶¹ but with practical value diminished since negative-going spikes are allowed. In 2010, Vogelstein and colleagues provided a fast non-negative deconvolution method that is, in addition to imposing a non-negative constraint on the spike trains, scalable on a large population of neurons³². Since our imaging involves hundreds of neurons over multiple contexts and days, we use the algorithm from Vogelstein *et al.* to deconvolve fluorescence signals.

Three classes of parameters were optimized in this algorithm to fit data: (1) GCaMP-related parameters, namely sensitivity of fluorescence to elevations in intracellular Ca^{2+} concentration (α) and baseline concentration (β); (2) acquisition parameters, namely the size of the time bin (δ) and the noise (σ) in the $\Delta F/F$ signals; and (3) system (hippocampus/CA3)-related parameters, namely expected spike rate per second (ϕ) and the time constant (τ), or the length it takes for Ca^{2+} concentrations to decay. For the GCaMP-related parameters, β is set to the baseline of the $\Delta F/F$ traces as described earlier, and α is set to 1 as a default value (since varying α broadly around this value did not affect the deconvolution results). δ was set to 1/3 s because image acquisition was at least 3 Hz per optical slice, and σ was estimated to be 0.16 as explained earlier. The main challenge resides in choosing parameters for ϕ and τ since (1) the expected spike rate (close to 0.1 Hz on average, but >10 Hz when bursting) is bimodal and insufficiently captured by the Poisson distribution of spikes as assumed by this model, and (2) the time constant expected for Ca^{2+} signals in hippocampus is not fully understood. Therefore, we optimized these two parameters by iterating over multiple combinations of time constants and expected spike rates to yield spike events consistent with good fits to our data (Extended Data Fig. 8). The final parameters chosen were: $\alpha = 1$; β = baseline; $\delta = 0.33$ s; $\sigma = 0.16$; $\phi = 5$ Hz; $\tau = 2$ s. These values were not exactly the same as, but were comparable to, values reported by others in cortical regions^{49,55,62}. Importantly, varying ϕ and τ within a fairly broad range ($\phi \sim 5$ –10 and $\tau \sim 0.67$ –2) did not significantly alter the main conclusions of the subsequent analyses. The $\Delta F/F$ signals for all the mice, contexts and days were deconvolved. Correlation coefficients were calculated on the deconvolved signals, and metrics that rely on accurate estimate of event onsets were recomputed, such as synchrony, lead-lag, and identification of HC neurons. The only difference from the methods described earlier was that there were no additional noise filters since the noise is filtered in the process of finding the optimal spike rate (here, event rate), and the onset time was characterized by the first instance that the signal became non-zero. Further analysis of the various specific deconvolution parameters would be of interest but would probably require combined

in vivo imaging and single-cell patching experiments, beyond the scope of the current study, and unlikely to significantly affect the specific analyses applied here given the robustness of results to broad ranges of parameters. Furthermore, we performed the analyses described earlier only to help ensure robustness in results obtained from using the raw $\Delta F/F$ for measurements relying on precise timing (correlations, leading versus lagging, and synchrony).

Virtual reality behavioural analysis. Lick rates and movements on the ball were captured in XML log files storing timestamps of behavioural data. These were then parsed with custom Python scripts and imported into MATLAB for synchronizing with microscope imaging frames with kHz precision, and for subsequent analysis. To quantify differences in licking between fear and neutral contexts during retrieval, the number of licks per second (each lick causing a beam-break resulting in TTL pulse output of at least 1V), was integrated over the first 2 min in the context. Total licking amounts were normalized to the highest lick rate, observed from any mouse in any context, and presented as a fraction of this value for each mouse and each context. Lick suppression data are presented as mean values across all mice in each experimental group; significance values of differences between contexts were evaluated by Student's *t*-test. Lick rates during optogenetic stimulation experiments were scored by quantification during the 15 s of light delivery, which was then normalized to the corresponding value from the 15 s just before light delivery. Significant differences in licking for fear versus neutral context, and for neutral/stimulated versus neutral context alone, were evaluated using Student's *t*-test. Lick rates and velocity on ball during synchronous population activity events were calculated by comparing the amount of licking and distance travelled in the 5 s window beginning at the start of a synchronous event, and then normalizing to the amount of licking and distance travelled in the 5 s window before synchronous event. Similar quantitative results were observed with this time window set to 1–10 s after synchrony compared with before, with no significant difference in lick rate and velocity during versus before synchrony.

Simultaneous 1P *in vivo* stimulation and 2P *in vivo* imaging. Simultaneous 1-photon (1P) stimulation (594 nm) and 2-photon (2P) imaging (920 nm) was performed by injecting the new red-shifted opsin with improved trafficking and kinetics (bReaChES) via AAV8 in cingulate and GCaMP6m via AAVdj in CA3, and positioning a cranial window above CA2/CA3 for optical access. The 2P imaging and full-field optogenetic stimulation setup is shown in Fig. 4b. Briefly, a resonant galvanometer 2P microscope using an NIR pulsed laser set to 920 nm is combined with simultaneous, full field stimulation using a 594 nm continuous wave laser that is coupled into the system with an optical fibre, lenses and dichroic beam splitters. A 2P compatible NIR reflecting dichroic designed with an additional 594 nm band-pass filter was used for 1P yellow light stimulation during 2P imaging. GCaMP6m signals (green channel) and stimulation artefact (red channel—used to precisely blank stimulation time points) are recorded using standard 2P resonant scanning imaging. 1P stimulation artefacts were removed offline from the 2P images. Stimulation parameters: 591 nm light, 20 Hz, 15 ms pulses, 15 s, 8–10 mW mm⁻² laser power at sample after the objective. In total, 4 mice (separate cohort from those used in the imaging-only experiments) were used for the combined stimulation and imaging experiments. The same cells and the same FOV are captured for before-training stimulation trials as well as after-training stimulation trials (conducted 5–7 days later). For a neuron to be considered responsive to (recruited by) the stimulus, at least one significant transient as defined earlier was required to occur during the stimulation window. For latency measurements provided in Fig. 4, event onsets were defined as the first time frame at which the response surpassed three standard deviations above noise, and increased for at least two consecutive frames; if occurring within the first frame, then only neurons with responses increasing from the previous frame are considered, to exclude responses decaying into the stimulation window. Responding neurons were assigned to latency bins of 333 ms.

DSI electrophysiology. DSI is dependent on the increase of postsynaptic intracellular Ca^{2+} to suppress GABA release from presynaptic inhibitory neurons expressing cannabinoid receptors⁶³. We performed patch-clamp recordings from CA3 neurons expressing GCaMP6m and examined spontaneous inhibitory postsynaptic currents (sIPSCs) before and following a depolarizing pulse to induce Ca^{2+} influx. Electrophysiological recordings were performed 4–6 weeks post-injection of AAVdj-CaMKII α ::GCaMP6m into CA3 (in 4–5-week-old mice). Coronal slices (300 μ m) from injected mice were prepared after intracardial perfusion with ice-cold, sucrose-containing artificial cerebrospinal fluid solution (ACSF; in mM): 85 NaCl, 75 sucrose, 2.5 KCl, 25 glucose, 1.25 NaH_2PO_4 , 4 MgCl_2 , 0.5 CaCl_2 and 24 NaHCO_3 . Slices recovered for 1 h at 32–34 °C, and then were transferred to an oxygenated recording ACSF solution (in mM): 123 NaCl, 3 KCl, 26 NaHCO_3 , 2 CaCl_2 , 1 MgCl_2 , 1.25 NaH_2PO_4 and 11 glucose, at room temperature. Excitatory synaptic transmission blockers (D-2-amino-5-phosphonovaleic acid (APV; 25 μ M) and 2,3-dihydroxy-6-nitro-7-sulfamoyl-benzo[f]-quinoxaline-2,3-dione (NBQX; 10 μ M)) were added to isolate GABAergic

postsynaptic currents, and 5 μ M carbachol was used to enhance sIPSC frequency to facilitate detection of DSI. Recordings were performed at 32–34 °C under constant perfusion of the oxygenated recording ACSF solution. Slices were visualized with an upright microscope (BX61WI, Olympus) under infrared differential interference contrast (IR-DIC) optics, and a Spectra X Light engine (Lumencor) was used for viewing GCaMP6m expression. Recordings of CA3 neurons were made after identifying GCaMP6m expression, and functional GCaMP6m activity was verifiable (in cells without BAPTA), by observing the increase in GCaMP6m fluorescence during the depolarizing pulse used to induce DSI. The following intracellular solution was used for the patch-clamp electrodes (in mM): 40 CsCl, 90 K-Gluconate, 1.8 NaCl, 1.7 MgCl₂, 3.5 KCl, 10 HEPES, 2 MgATP, 0.4 Na₂GTP, 10 phosphocreatine (pH 7.2, 270–290 mOsm). For the BAPTA experiments, 40 mM BAPTA was added to the intracellular solution. Series resistance was monitored for stability, and recordings were discarded if the series resistance changed significantly (by >20%) or reached 20 M Ω . Resting membrane potential was taken at rest, and the reported values incorporate a liquid junction potential of +11.2 mV. Input resistance was calculated from a 100 pA pulse. MiniAnalysis (Synaptosoft) and pClamp10.3 (Molecular Devices) was used to calculate charge transfer (area under sIPSCs) and analyse data. Baseline charge transfer was measured during a 4 s pre-pulse period, DSI was examined during a 4 s period following the depolarizing pulse, and charge transfer after recovery from DSI was measured during a 4 s window. The pulse used to evoke DSI was a 500 ms step to 0 mV from holding potential of –65 mV.

bReaChES design and testing

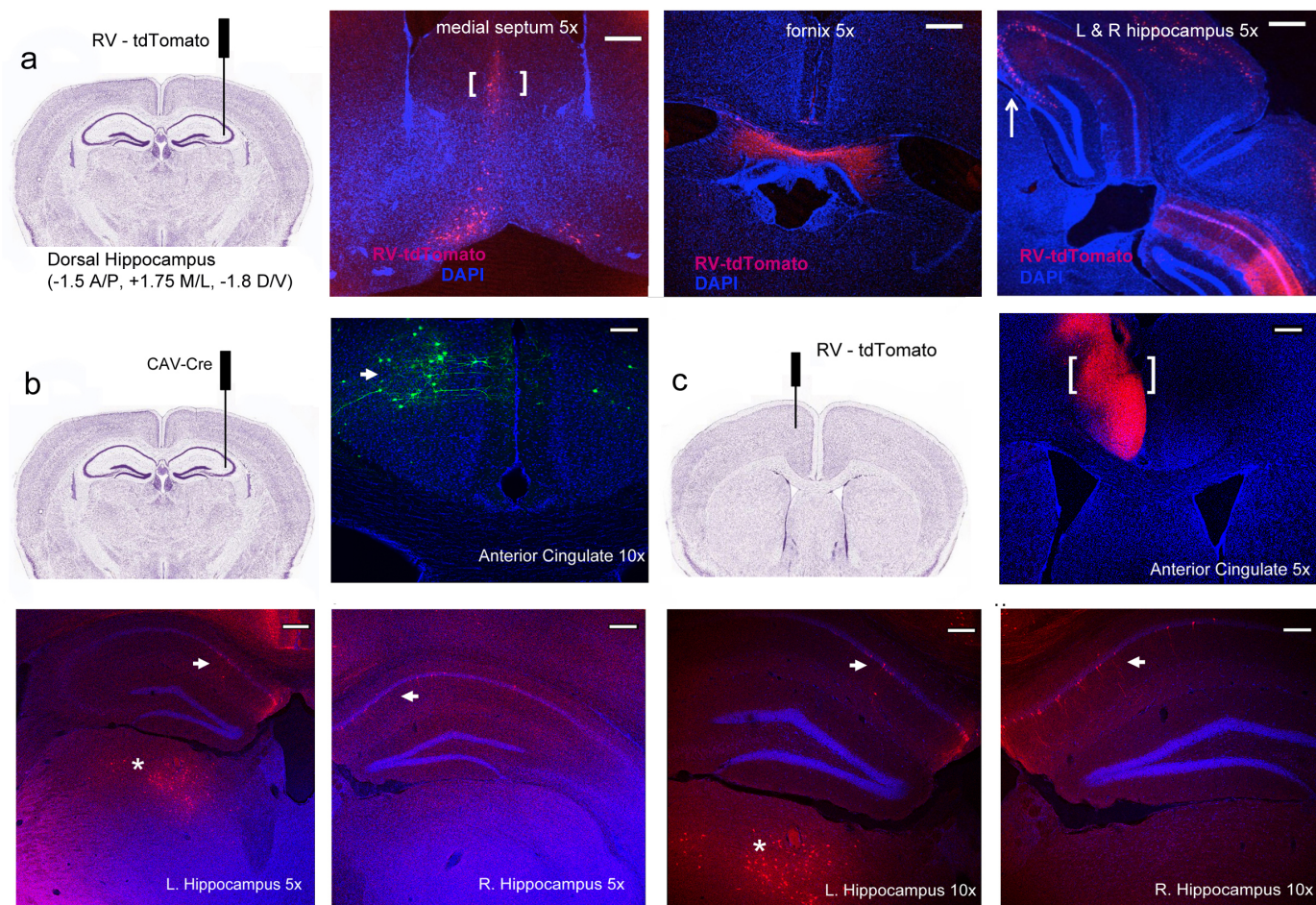
Cloning and single mutagenesis of ReaChR and bReaChES. DNA sequences of ReaChR and bReaCh were synthesized (GenScript) and cloned into AAV vectors containing the CamKII α promoter for expression in neurons. All constructs were fused to eYFP DNA to detect protein expression in neurons by fluorescence microscopy. The Glu123Ser mutation was introduced using QuickChange Site-Directed mutagenesis kit (Agilent). Plasmid DNA was purified with QIAprep Spin Miniprep Kits (Qiagen) after transformation and amplification in *Escherichia coli*.

Electrophysiological recordings in cultured hippocampal neurons. Electrophysiological recordings in neuronal cultures were prepared as described⁶⁴. Patch pipettes (4–6 M Ω) were pulled from glass capillaries (Sutter Instruments) with a horizontal puller (P-2000, Sutter Instruments) for whole-cell recordings in voltage and current clamp. Recordings were made using a MultiClamp700B amplifier (Molecular Devices). The external recording solution contained (in mM): 127 NaCl, 10 KCl, 10 HEPES, 2 CaCl₂, 2 MgCl₂, 30 D-glucose, pH 7.3, including synaptic blockers (25 μ M D-APV, 10 μ M NBQX). The patch pipette solution contained (in mM): 140 K-gluconate, 10 HEPES, 10 EGTA, 2 MgCl₂, pH 7.3. All measurements were corrected for a liquid junction potential of +15 mV. Series resistance was monitored throughout recordings for stability. A Spectra X Light engine (Lumencor) was used to excite eYFP and to apply light for opsin activation. Yellow and red stimulation light was filtered by 575/25 or 632/22 band-pass filters (Chroma) and applied through a $\times 40$ objective (Olympus) at 5 mW mm^{–2} light intensity. Light power density was measured with a power meter (Thorlabs). The functionality of all constructs was determined by comparing stationary photocurrents at –80 mV to 1 s continuous light pulse. Spikes were optically evoked in current-clamp mode with light pulses (5 ms) delivered at 633 nm, 5 mW mm^{–2} and 1–20 Hz. The activation spectra for C1V1, bReaCh-ES and ChR2 was measured by recording stationary photocurrent in voltage clamp mode at –80 mV and light intensities of 0.65 mW mm^{–2} at each wavelength. Light was delivered through 20 nm band-pass filters (Thorlabs) at (in nm): 400, 420, 440, 460, 470, 480, 490, 500, 520, 540, 560, 570, 580, 590, 600, 620, 630, 650. Photocurrents were normalized to maximum values respectively: 480 nm for ChR2, 560 nm for C1V1 and 570 nm for bReaCh-ES. Kinetics of channel closure were determined by fitting the decay of photocurrents after light off, with mono-exponential functions. Channel kinetics were quantified by corresponding τ_{off} values respectively. pClamp10.3 (Molecular Devices) and OriginLab8 (OriginLab) software was used to record and analyse data.

Stereotactic virus injection of bReaCh-ES. The following adeno-associated viruses (AAVs) with serotype DJ were produced at the Stanford Neuroscience Gene Vector and Virus Core: AAVDJ-CaMKII::bReaCh-E162S-TS-eYFP; and AAVDJ-CaMKII::C1V1(E122T/E162T)-TS-eYFP. Four-to-six-week-old mice were injected bilaterally with 1 μ l of either virus in the medial prefrontal cortex, at the following coordinates (from Bregma): A/P: +1.7 mm; M/L: +0.3 mm; D/V: –2.5 mm. Titre was matched at 1.5×10^{12} vg ml^{–1} for both viruses.

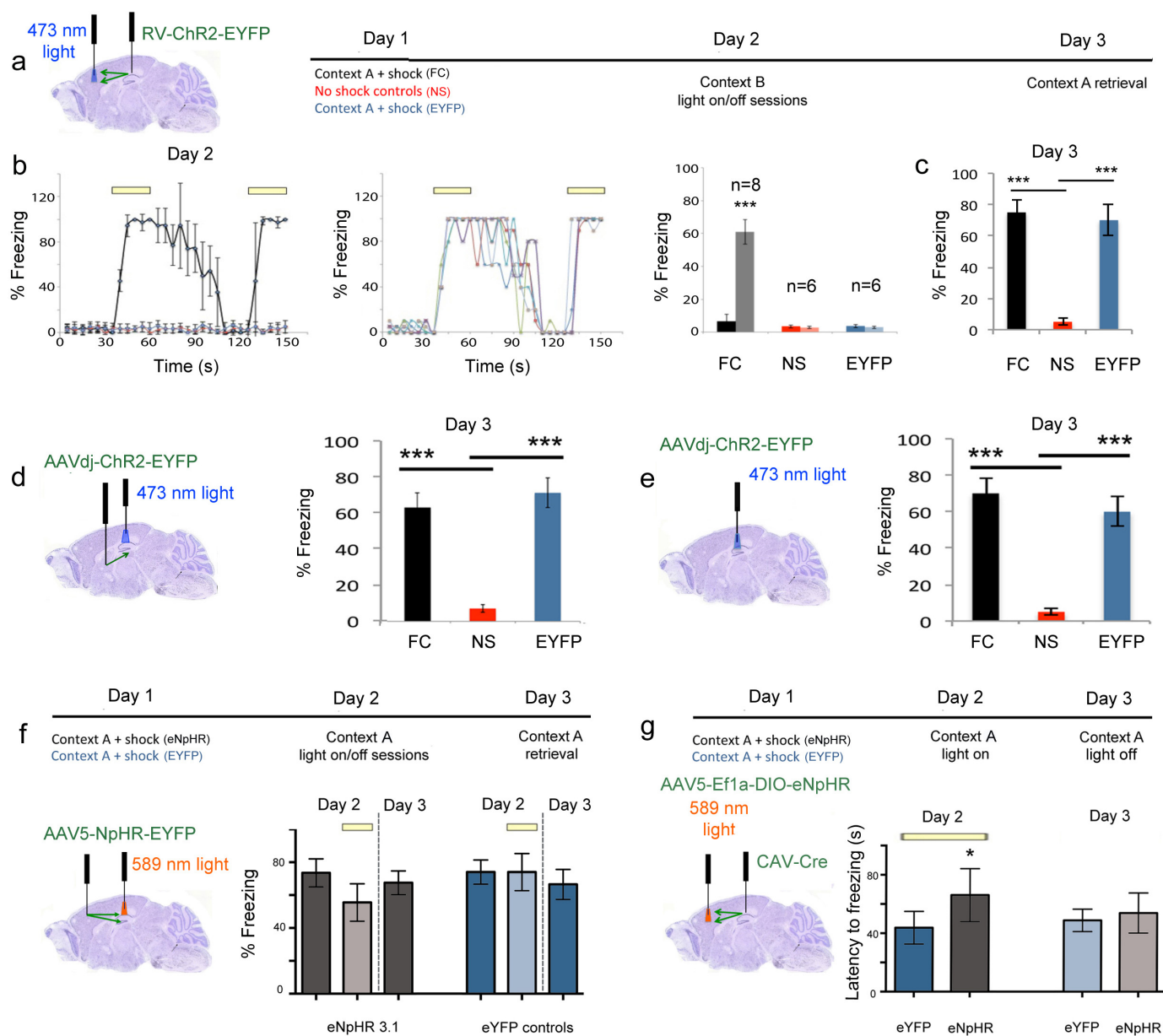
Slice electrophysiology for bReaCh-ES characterization in mPFC and BLA. Electrophysiological recordings were performed 12–14 weeks post-injection for opsin expression at the mPFC terminals. Coronal slices (300 μ m) from injected mice were prepared after intracardial perfusion with ice-cold, sucrose-containing ACSF (in mM): 85 NaCl, 75 sucrose, 2.5 KCl, 25 glucose, 1.25 NaH₂PO₄, 4 MgCl₂, 0.5 CaCl₂ and 24 NaHCO₃. Slices recovered for 1 h at 32–34 °C, and then were transferred to an oxygenated recording ACSF solution (in mM): 123 NaCl, 3 KCl, 26 NaHCO₃, 2 CaCl₂, 1 MgCl₂, 1.25 NaH₂PO₄ and 11 glucose, at room temperature. Electrophysiological recordings were performed at 32–34 °C under constant perfusion of the oxygenated recording ACSF solution. For mPFC recordings, synaptic transmission blockers (APV (25 μ M), NBQX (10 μ M) and gabazine (10 μ M)) were added to the recording ACSF solution. Slices were visualized with an upright microscope (BX61WI, Olympus) under IR-DIC optics. A Spectra X Light engine (Lumencor) was used both for viewing fluorescent protein expression and delivering light pulses for opsin activation. Light power density was obtained with a power meter (Thorlabs). Recordings of mPFC neurons were made after first identifying regions of eYFP⁺ expression, and recordings of postsynaptic basolateral amygdala (BLA) neurons were obtained after confirming eYFP⁺ expression in both mPFC and the mPFC axonal fibres at the BLA. Whole-cell voltage-clamp recordings were performed at –65 mV, and current-clamp recordings were performed at rest. Patch-clamp pipettes contained the following internal solution (in mM): 125 K-gluconate, 10 KCl, 10 HEPES, 4 Mg₃-ATP, 0.3 Na-GTP, 10 phosphocreatine, 1 EGTA. Recordings were conducted using MultiClamp700B amplifier and pClamp10.3 software (Molecular Devices). pClamp10.3, OriginLab8 (OriginLab), and SigmaPlot (SPSS) were used to analyse data. Stationary photocurrent of the opsins was measured at the end of a 1 s light pulse in voltage-clamp mode. Light-evoked spike probability in the mPFC neurons and in the postsynaptic BLA neurons was calculated as the fraction of successful action potentials evoked in the recorded neuron upon various light stimulation frequencies. Light-evoked EPSC amplitude in the postsynaptic BLA neurons was measured at the peak of the evoked response to light stimulation of the opsin-expressing mPFC fibres. Series resistance was monitored for stability, and recordings were discarded if series resistance changed significantly (by >20%) or reached 20 M Ω . Statistical analysis was performed with two-tailed *t*-test, with significance set at *P* < 0.05.

- Harvey, C. D., Coen, P. & Tank, D. W. Choice-specific sequences in parietal cortex during a virtual-navigation decision task. *Nature* **484**, 62–68 (2012).
- Niell, C. M. & Stryker, M. P. Modulation of visual responses by behavioral state in mouse visual cortex. *Neuron* **65**, 472–479 (2010).
- Mukamel, E. A., Nimmerjahn, A. & Schnitzer, M. J. Automated analysis of cellular signals from large-scale calcium imaging data. *Neuron* **63**, 747–760 (2009).
- Canny, J. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **8**, 679–698 (1986).
- Ko, H. et al. Functional specificity of local synaptic connections in neocortical networks. *Nature* **473**, 87–91 (2011).
- Cossell, L. et al. Functional organization of excitatory synaptic strength in primary visual cortex. *Nature* **518**, 399–403 (2015).
- Tarjan, R. E. Depth first search and linear graph algorithms. *SIAM J. Comput.* **1**, 146–160 (1972).
- Sedgewick, R. *Algorithms in C++, Part 5 Graph Algorithms* (Addison-Wesley, 2002).
- Schwartz, T. H. et al. Networks of coactive neurons in developing layer 1. *Neuron* **20**, 541–552 (1998).
- Greenberg, D. S., Houweling, A. R. & Kerr, J. N. Population imaging of ongoing neuronal activity in the visual cortex of awake rats. *Nature Neurosci.* **11**, 749–751 (2008).
- Holekamp, T. F., Turaga, D. & Holy, T. E. Fast three-dimensional fluorescence imaging of activity in neural populations by objective-coupled planar illumination microscopy. *Neuron* **57**, 661–672 (2008).
- Smith, S. L. & Häusser, M. Parallel processing of visual space by neighboring neurons in mouse visual cortex. *Nature Neurosci.* **13**, 1144–1149 (2010).
- Castillo, P. E., Younts, T. J., Chávez, A. E. & Hashimoto, Y. Endocannabinoid signaling and synaptic function. *Neuron* **76**, 70–81 (2012).
- Berndt, A., Lee, S. Y., Ramakrishnan, C. & Deisseroth, K. Structure-guided transformation of channelrhodopsin into a light-activated chloride channel. *Science* **344**, 420–424 (2014).
- Lee, S. Y., Földy, C., Szabadics, J. & Soltesz, I. Cell-type-specific CCK2 receptor signaling underlies the cholecystokinin-mediated selective excitation of hippocampal parvalbumin-positive fast-spiking basket cells. *J. Neurosci.* **31**, 10993–11002 (2011).
- Kohara, K. et al. Cell type-specific genetic and optogenetic tools reveal hippocampal CA2 circuits. *Nature Neurosci.* **17**, 269–279 (2014).
- Varga, C., Lee, S. Y. & Soltesz, I. Target-selective GABAergic control of entorhinal cortex output. *Nature Neurosci.* **13**, 822–824 (2010).



Extended Data Figure 1 | Anatomical characterization of the AC-CA projection. **a**, Five days after RV-tdT injection into the hippocampus (coordinates specified), retrogradely labelled neurons were detected in the contralateral hippocampus (arrow), medial septum (bracket), and AC (Fig. 1a). Scale bars: $\times 5$: 300 μm ; $\times 10$: 100 μm (confocal). **b**, Eight weeks after injection of CAV-Cre into the hippocampus and DIO-eYFP in the AC (coordinates specified), afferent cell bodies were detected in the AC (arrow). Confocal image;

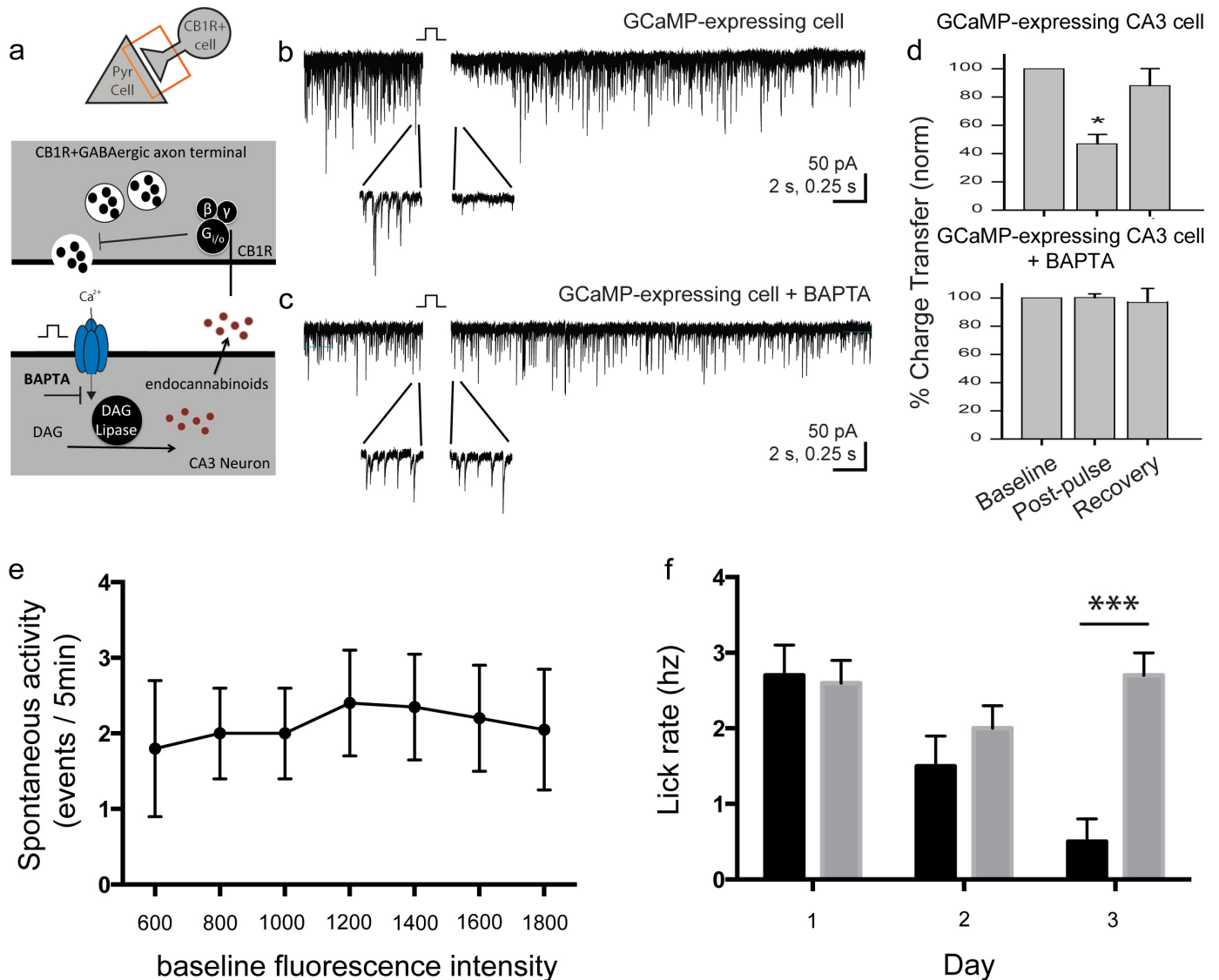
$\times 10$ magnification; scale bar: 100 μm . **c**, Retrograde tracing with RV-tdT from the AC to map reciprocal connections from the hippocampus. The injection site is indicated in brackets, sparse labelling of afferent cell bodies in left and right hippocampus, primarily in the subiculum (arrows), and also in the medial dorsal thalamic nucleus as expected (asterisk). Confocal $\times 5$, scale bar, 200 μm ; $\times 10$ images, scale bar 100 μm .



Extended Data Figure 2 | Optogenetic manipulation of the AC–CA

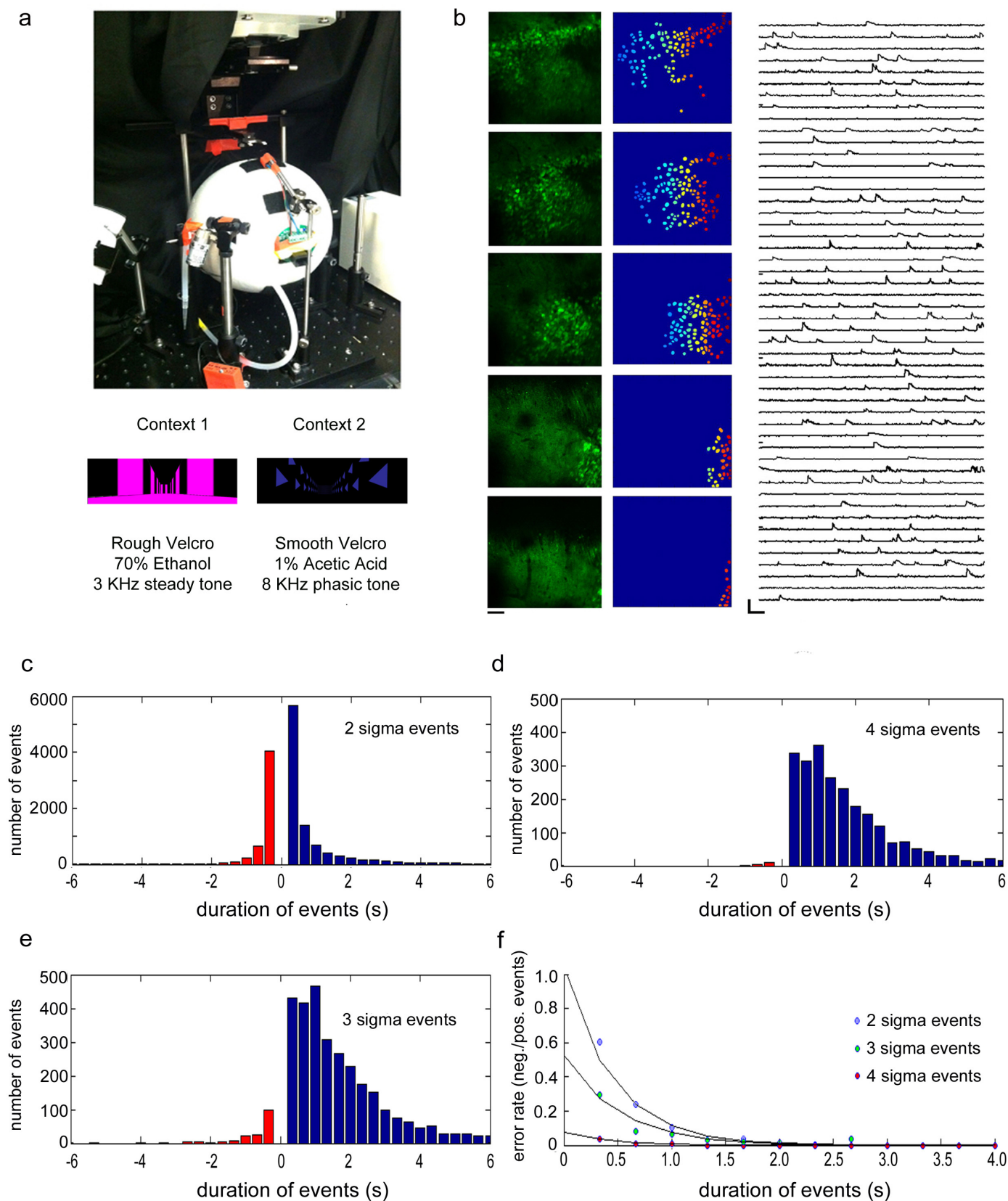
projection. **a**, Experimental design: RV-ChR2-eYFP (or eYFP alone) was injected into the dorsal hippocampus and light was delivered above the cell bodies in the AC. Five days after injection, ChR2 and eYFP mice were fear conditioned in context A while no-shock controls were only exposed to context A (day 1). All mice were tested with light on and off sessions in context B (day 2), and then tested for contextual memory retrieval in context A (day 3). Optogenetic stimulation was with 473 nm light in a train of 20 Hz, 15 ms pulses, 30 s duration, with 8–10 mW laser power at fibre tip. **b**, Freezing (no head motion observed) during day 2 is plotted in 5 s time bins over 150 s in context B (left). ChR2/shock (FC): black; ChR2/no shock (NS): red; eYFP/shock (eYFP): blue. Individuals in the FC group (each animal a different colour) are shown (middle). Summary (right): percentage time freezing (mean \pm s.d.) 20 s before light on (darker shade) versus 20 s after light on (lighter shade); FC: $60.9 \pm 7.4\%$ light on versus $6.5 \pm 4.4\%$ light off, $n = 8$; NS: $2.7 \pm 0.65\%$ light on versus $3.4 \pm 0.95\%$ light off, $n = 6$; eYFP: $2.9 \pm 0.75\%$ light on versus $3.6 \pm 1\%$ light off, $n = 6$; $P < 0.001$, two-way ANOVA with repeated measures. **c**, Preservation of contextual fear memory (percentage time freezing) on day 3 in the original context (mean \pm s.d., $P < 0.001$, unpaired t -test). **d**, Preservation of contextual memory in medial septum injected mice (Fig. 1f); percentage time freezing on day 3 in the original context (mean \pm s.d., $P < 0.001$, comparisons shown, unpaired t -test). **e**, Preservation of contextual memory in hippocampus injected mice (Fig. 1h); percentage time freezing on day 3 in the original context (mean \pm s.d., $n = 8$ mice, $P < 0.001$, paired t -test). **f**, The successful loss-of-function experiments targeting hippocampus-dependent memory formation mediated by cells giving rise to the AC–CA projection (reported in Fig. 2) were designed to allow the most robust inhibition of this circuit element. An alternative design (attempting to target the projection field despite the broad and long septotemporal extent of the hippocampal

formation) was also explored as shown here but was not effective, as expected; we injected AAV5-eNpHR3.0-eYFP (or AAV5-eYFP in a parallel cohort) bilaterally into the AC, and targeted light stimulation bilaterally to axon terminals in the hippocampus. Eight weeks after injection, all mice were fear conditioned to context (day 1), and tested for context retrieval during light on/off sessions (day 2), and again for context retrieval in light-off only (day 3). Optogenetic inhibition was with constant illumination of 589 nm light, 30 s duration, with 8–10 mW laser power at fibre tip. **g**, We observed a trend towards reduction in freezing due to optical inhibition of the AC–CA projection during memory retrieval. Percentage time freezing in context A during day 2 before light (darker bar on left) versus after light (lighter bar at right). eNpHR3.0: $73.5 \pm 8.5\%$ light off versus $55.5 \pm 11.4\%$ light on, $n = 10$; eYFP: $74 \pm 7.4\%$ light off versus $74 \pm 11.3\%$ light on, $n = 10$; percentage time freezing in context A with light off (dark bars) during day 3 is shown after dotted line. eNpHR3.0: $67.5 \pm 7.2\%$, $n = 10$; eYFP: $66.5 \pm 9.1\%$, $n = 10$ ($P = 0.067$, two-way ANOVA). As expected, point illumination may be less effective for inhibiting broad axon terminal field volumes. **h**, Extension of findings from effective loss-of-function experiments (Fig. 2) targeting hippocampus-dependent memory formation mediated by cells giving rise to the AC–CA projection: significant effect on speed of onset of memory expression. Experimental design: CAV-Cre was injected into the dorsal hippocampus, DIO-eNpHR3.0 (or DIO-eYFP) was injected into the AC, and light was delivered above cell bodies in the AC. All mice were fear conditioned in context A (day 1), tested for latency to contextual retrieval with light-on only (day 2), and then for latency to context retrieval in light-off only (day 3). **i**, Day 2: 66.1 ± 18.1 s for eNpHR3.0 ($n = 12$) versus 43.8 ± 11.1 s for eYFP ($n = 8$) during light on; day 3: 53.8 ± 13.7 s for eNpHR3.0 versus 48.8 ± 7.7 s for eYFP during light off; $P < 0.05$ two-way ANOVA with repeated measures. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.



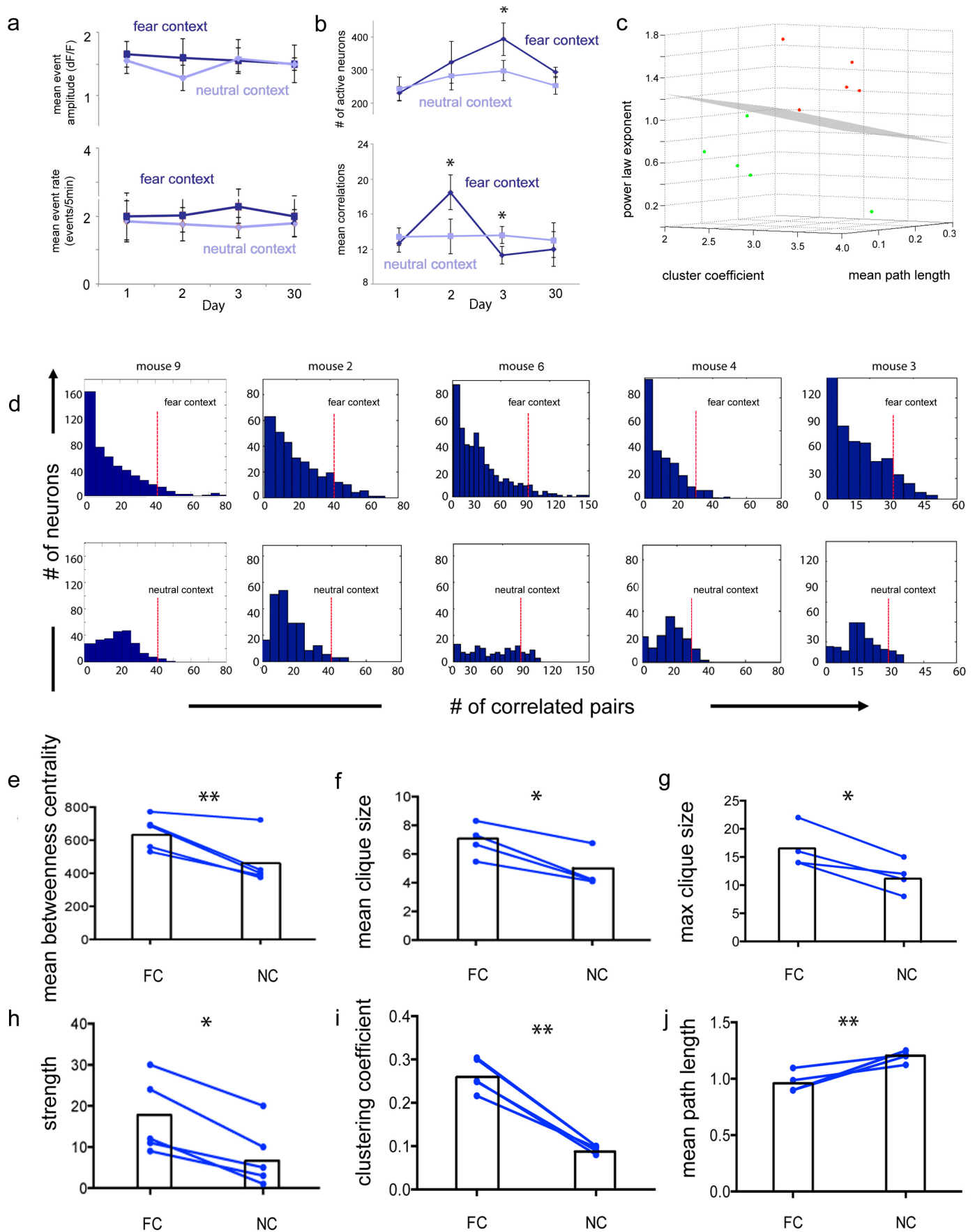
Extended Data Figure 3 | Physiological properties of GCaMP6m-expressing CA3 neurons. **a**, To ensure that expression of GCaMP6m did not alter Ca^{2+} -related physiological processes, we tracked a form of endocannabinoid-mediated short-term plasticity known as depolarization-induced suppression of inhibition (DSI). Schematic diagram of DSI is shown; DSI is dependent on the increase of postsynaptic intracellular Ca^{2+} to trigger the synthesis and release of endocannabinoids, which then signal in a retrograde fashion to suppress GABA release from presynaptic inhibitory neurons expressing cannabinoid receptors (adapted with permission from ref. 65). Intrinsic membrane properties of the GCaMP6m-expressing CA3 cells were similar to previously reported values for CA3 (ref. 66); mean resting potential: -72.1 ± 1.6 mV; mean input resistance: 161.8 ± 26.4 M Ω ; $n = 7$. **b**, Sample trace illustrating DSI of sIPSCs in a GCaMP6m-expressing CA3 cell following application of a depolarizing current step (from -65 mV to 0 mV for 500 ms). **c**, Sample trace illustrating lack of DSI of sIPSCs with inclusion of the intracellular calcium chelator BAPTA ($1,2$ -bis(o -aminophenoxy)ethane- N,N,N',N' -tetraacetic acid) in the patch pipette. **d**, Summary graph of normalized charge transfer in GCaMP6m-expressing cells with standard intracellular solution (left, normalized charge transfer of sIPSCs following DSI

compared with pre-pulse baseline over the same fixed interval: charge reduced to $46.9 \pm 6.7\%$ of baseline charge; $n = 7$; comparable to charge transfer reported for non-GCaMP-expressing cells⁶⁷) and with addition of intracellular BAPTA (right, $n = 6$; error bars represent standard error of the mean (s.e.m.); $P < 0.05$, paired t -test). **e**, Spontaneous event rate (detection described in Methods) of GCaMP6m-expressing neurons as a function of baseline GCaMP6m fluorescence intensity (arbitrary units spanning the range over which event-rate population data could be reliably quantified) in each cell (pooling all neurons with ≥ 1 significant transient, from all mice, over all fields of view). Event rates were not observed to change significantly as a function of GCaMP6m expression level (Spearman's rank correlation coefficient: 0.48 , $P = 0.3$). **f**, Behavioural scores from mice before GCaMP6m virus injection and implantation of cannulae above hippocampus; lick rates for the first 2 min in the fear (black) versus neutral (grey) contexts are provided. The level of learning assessed by lick suppression on day 3 retrieval (mean 0.5 ± 0.3 for day 3 fear versus 2.7 ± 0.3 for day 3 neutral; $n = 10$, $P < 0.001$, paired t -test) pre-injection/implantation was comparable to levels corresponding to post-injection/implantation (compare with Fig. 3b). $*P < 0.05$, $**P < 0.01$, $***P < 0.001$.



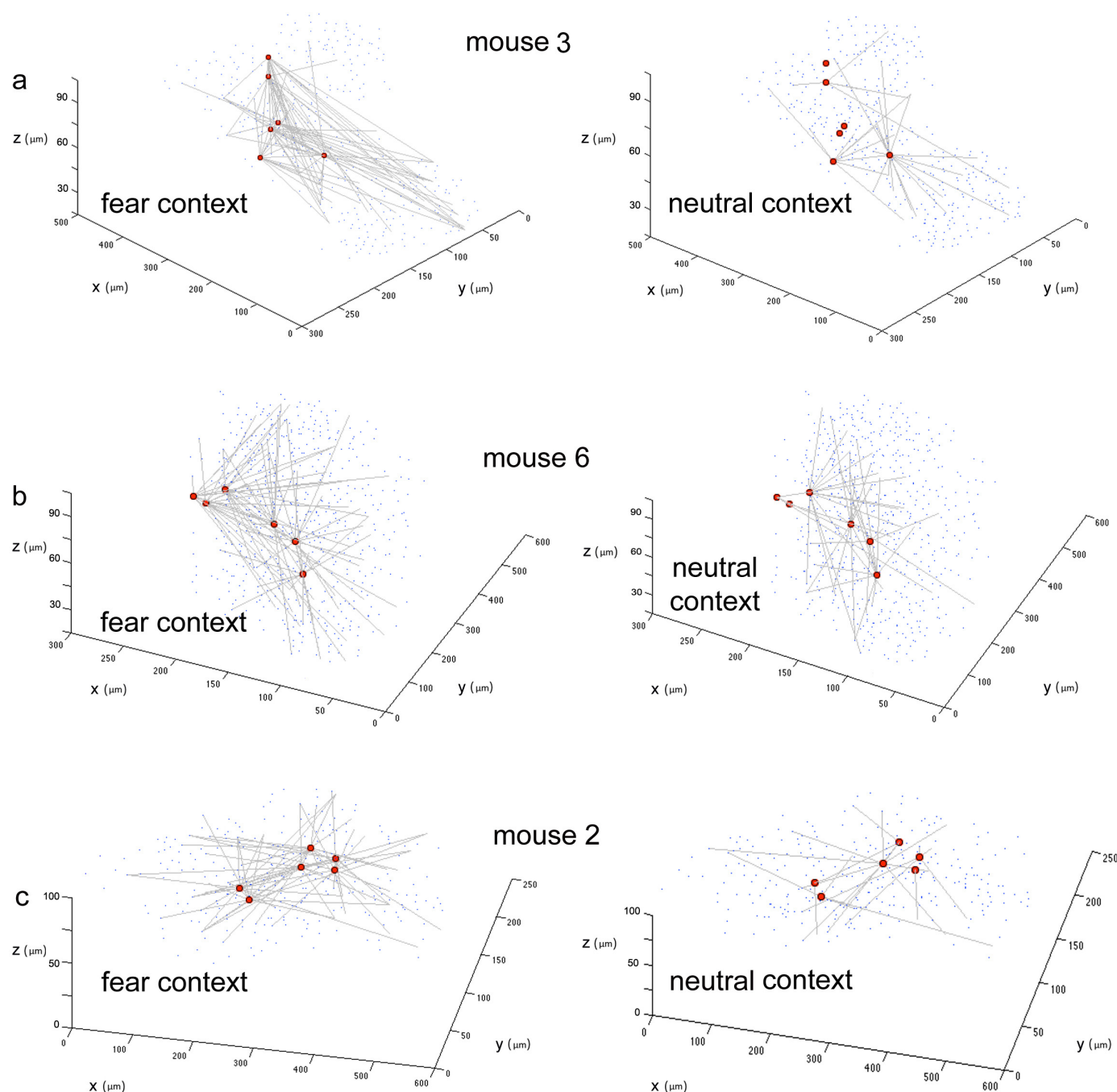
Extended Data Figure 4 | Real-time imaging of neural ensembles in three-dimensional hippocampal volumes: extraction of neural sources and identification of significant transients. **a**, Head-fixed virtual reality setup. Mice run on an axially fixed track ball³¹ while movements and licking behaviour are measured through an optical mouse and a lickometer, respectively, both interfaced with the virtual-reality gaming software. For contextual fear conditioning, water-restricted mice were exposed to two contexts with distinct visual, olfactory, tactile and auditory cues (day 1), and provided with aversive air puffs in one context (fear context), but not the other (neutral context) (day 2). Fear memory retrieval in the two contexts was quantified (days 3, 30) by lick suppression. **b**, Sample mean intensity z projections from raw videos (scale bar: 50 μm), with extracted neural sources (segmented cells) from CA3 for each of the optical sections, along with the first 50 time-series traces. Scale: 300% $\Delta F/F$, 30 s. **c**, Identification of significant transients in $\Delta F/F$

traces. Histogram showing the distribution of events occurring at amplitude 2σ above noise (noise calculated on a per-cell basis), over a range of event duration in seconds. The number of negative-going transients at each amplitude and duration are plotted in red to the left of the ordinate, and positive-going transients at each amplitude and duration are plotted in blue to the right. **d, e**, The above analysis is repeated for events that occur at an amplitude of 3σ (**d**) and 4σ (**e**). **f**, False positive rates for 2-, 3-, and 4- σ events (pooled across all neurons in all mice over all FOVs). False positive rate curves were calculated for each σ level by dividing the number of negative events at that level by the number of positive events at that level (Methods). Event onset was defined as the time corresponding to $\Delta F/F$ exceeding 2σ , and offset as the time corresponding to $\Delta F/F$ falling below 0.5σ . A decaying exponential was fit by least-squares to the false positive rate values, allowing for the determination of a minimum transient duration at each σ level for different confidence levels.



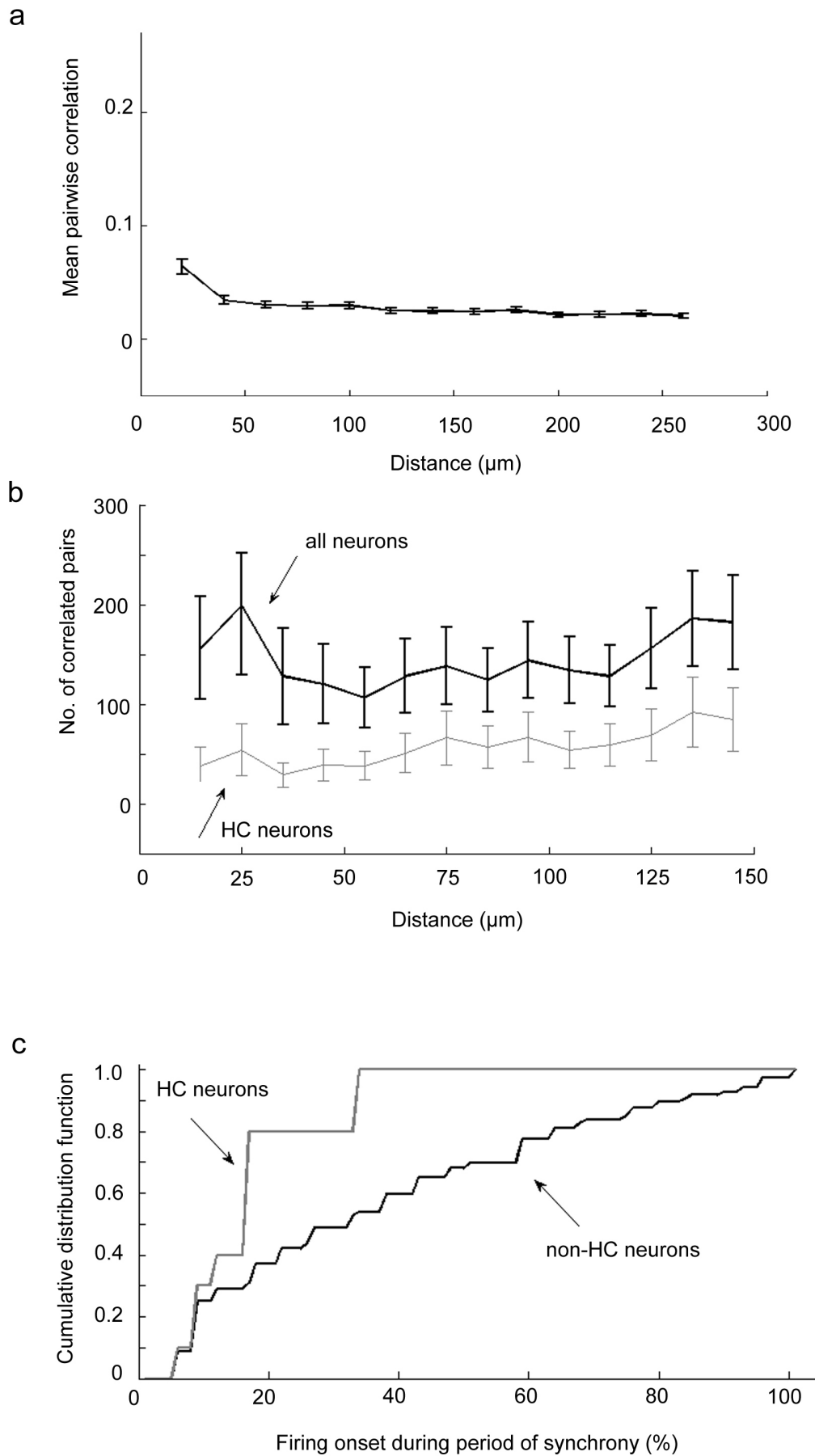
Extended Data Figure 5 | Cell populations and graph properties of fear and neutral networks in hippocampus during retrieval. **a**, No context-dependent change in total event amplitude or rate was detected. Top, mean GCaMP6m-detected event amplitude (average $\Delta F/F$ of all significant events; definition of significant event for each neuron as described in Methods) is plotted across days for mice in the fear and neutral contexts ($n = 5$ mice, not significant in paired t -tests). Bottom, mean GCaMP6m-detected event rate plotted across days for mice in the fear and neutral contexts ($n = 5$ mice, not significant in paired t -tests). **b**, Context-dependent changes in individual-neuron and correlated behaviour were observed. Top, number of active neurons (at least one significant GCaMP6m transient detected within first 2 min in context) plotted for fear and neutral contexts ($n = 5$ mice, 378 ± 64 for day 3 fear context versus 257 ± 39 for day 3 neutral context; $P < 0.05$, paired t -test, mean \pm s.d.). Bottom, mean number of correlated pairs per neuron (where pairwise Pearson's correlation coefficient > 0.3) plotted for fear and neutral contexts ($n = 5$ mice, 18.5 ± 1.8 for day 2 fear context versus 13.4 ± 1.4 for day 2 neutral context; 11.3 ± 0.8 for day 3 fear context versus 13.6 ± 0.5 for day 3 neutral context; $P < 0.05$, paired t -test). **c**, Fitting histograms from Fig. 3d to an exponential distribution of the form ae^{-bx} demonstrates a power-law ($b > 1$) distribution in day 3 fear context (each red dot represents one mouse) compared to day 3 neutral context ($b < 1$; green dots), which was consistent across all mice ($n = 5$ mice; $P < 0.01$, paired t -test). Many graph

properties were calculated for fear versus neutral context, but the power-law exponent of the degree distribution distinguished fear (red) from neutral (green) most powerfully (discriminants shown: coefficient of the power law exponent = 0.78, coefficient of cluster coefficient = 0.61, coefficient of mean path length = 0.11, with 90% confidence intervals being [0.74, 1.0], [0.1, 0.65] and [0.01, 0.23], respectively). These confidence intervals were obtained using 1,000 bootstrapped samples; shown is the best three-dimensional hyperplane separation using a linear support vector machine classifier. **d**, Histograms of the number of correlated partner neurons existing for each neuron in fear versus neutral context on day 3 (retrieval testing) across mice. The dotted red line indicates correlation threshold (set automatically as mean + 1 standard deviation in the number of correlated pairs in the network), to the right of which lie (by definition) the highly correlated or HC neurons. Similar measurements of interest in fear versus neutral context across mice were calculated and are provided here for other graph invariant properties: **e**, betweenness centrality; **f**, **g**, clique properties; **h**, strength; **i**, cluster coefficient; and **j**, mean path length (all defined in Methods). For the above calculations, correlation between two neurons was defined to exist when the pairwise Pearson correlation coefficient exceeded 0.3 (Methods). Data are presented as individual data points corresponding to each mouse, with mean \pm s.d. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, using paired t -tests.



Extended Data Figure 6 | Functional relationships of fear-context-defined HC neurons as appearing in fear versus neutral context. a–c, Data from all additional mice (beyond the exemplar of Fig. 3e, f) demonstrating that HC neurons (red circles) in the fear context with a high degree of correlated partners (grey edges) when located in the neutral context have a much lower degree of correlated partners ($n = 4$ mice including the example in Fig. 3e, f);

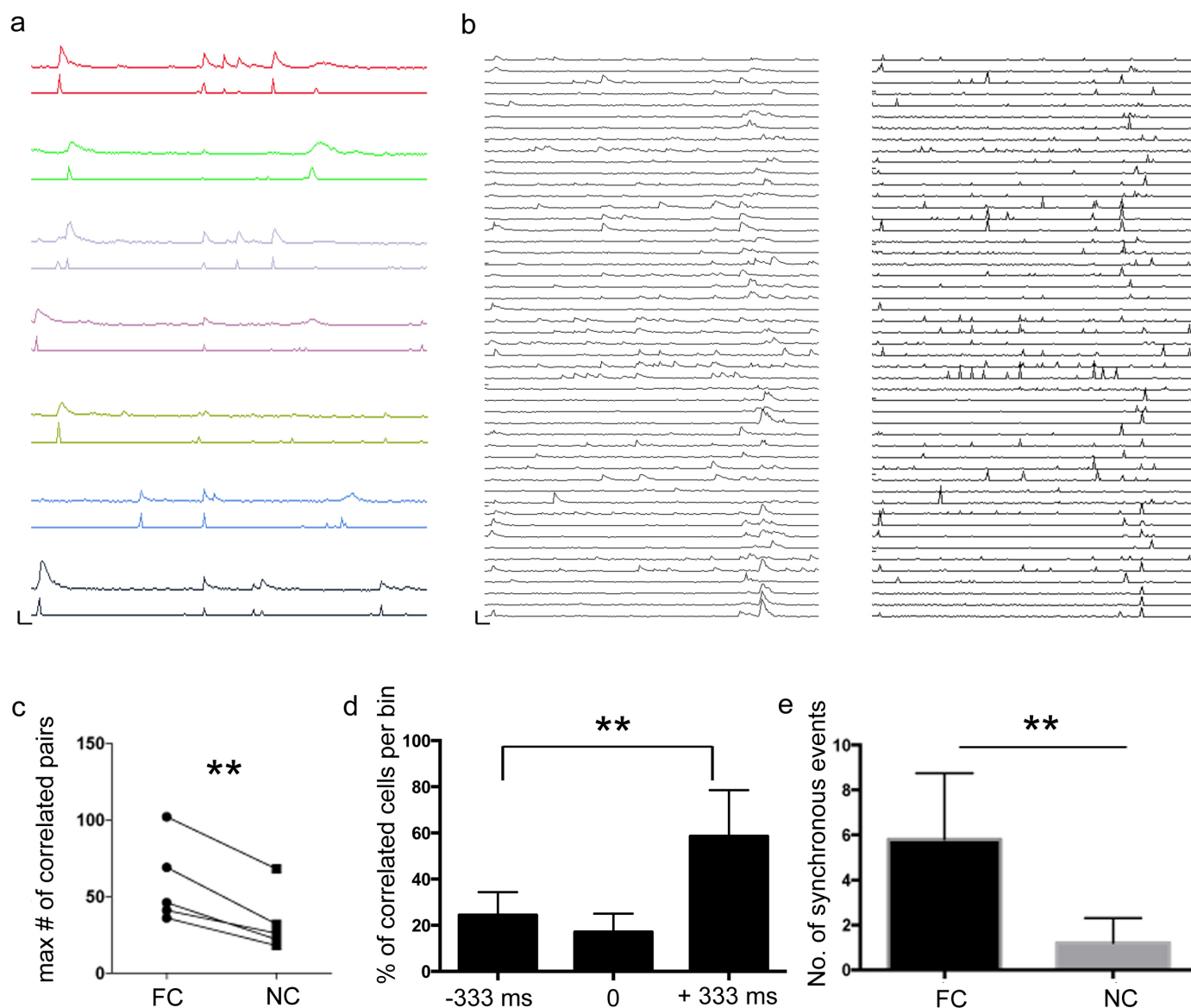
mean = 60 correlated pairs (standard deviation 19.4) in fear context versus mean = 18 correlated pairs (standard deviation 14.2) in neutral context; $P < 0.01$ by paired t -test). Only four mice are analysed here because the exact same FOV (with cell identities) was not captured in fear versus neutral context for one mouse.



Extended Data Figure 7 | Spatial and temporal organization of HC neurons.

a. Plot of mean pairwise correlation versus mean pairwise distance averaged over all FOVs (all days and all contexts) from all five mice. It was possible to detect a significant but weak relationship between mean correlation and distance (Spearman's correlation = -0.66 , $P = 0.01$), which could be a reflection of fine-scale spatial clustering as might be expected of recurrent circuits in CA3, but would also probably include residual crosstalk between regions of interest (ROIs) due to brain motion and common neuropil signal, which is expected and not significantly different from what has been previously observed in the hippocampus³. **b.** Plot of the number of correlated pairs versus pairwise distance for all neurons (black line), and HC neurons only (grey line). More correlated pairs were found at greater distances for HC

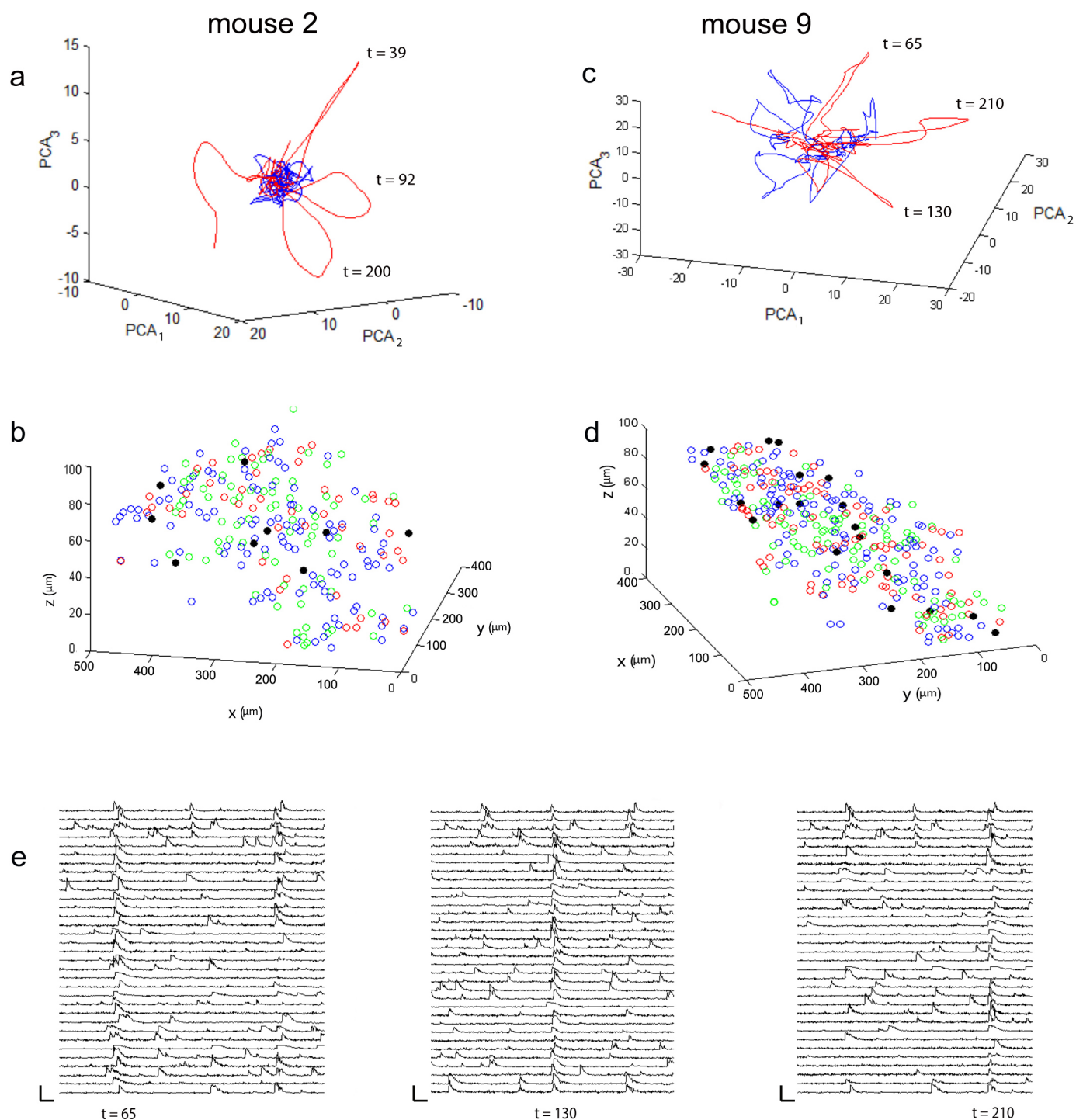
neurons (Spearman's = 0.84 , $P = 0.002$ for HC neurons; Spearman's correlation = 0.23 , $P = 0.43$ for all neurons). **c.** Cumulative distributions showing fraction of HC neurons (y -axis) with onset times at various latencies across the time course of synchronous events (x -axis) averaged across all mice, compared to response latencies of non-HC neurons. HC neuron activity appeared significantly earlier than for non-HC neurons during synchronous events ($P < 0.001$, Kolmogorov-Smirnov two-tail test, $\kappa = 0.664$; note the horizontal resolution of the plot is inversely proportional to length of the synchrony window, and dependent on frame duration; for instance, a 10-s-long synchrony window with frame duration of 333 ms corresponds to a 3.33% resolution per frame).



Extended Data Figure 8 | Additional analyses: estimation of event onsets using fast non-negative deconvolution, and correlated pair analysis.

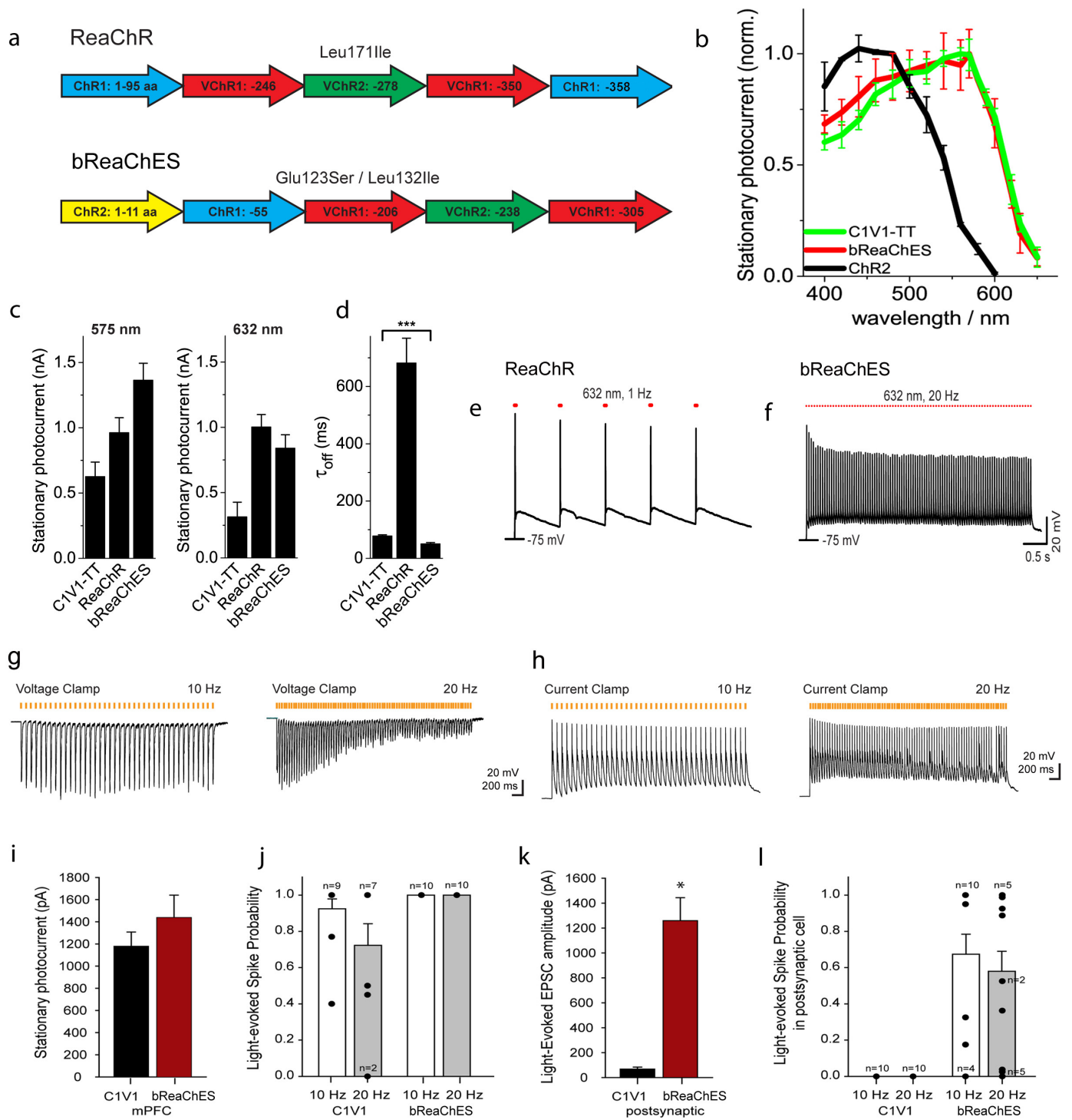
a, Example pairings of the original GCaMP6m trace (top traces), with the deconvolved trace (bottom traces), shows reliable estimation of event onset from deconvolved data (deconvolution algorithm and parameters detailed further in Methods). Scale bar: 150% $\Delta F/F$, 10 s. **b**, Original GCaMP6m traces from a representative synchronous event in one animal (left), paired with the deconvolved traces for that same synchronous event (right). Scale bar: 300% $\Delta F/F$, 10 s. **c**, The highest-degree node (neuron with the greatest number of correlated pairs) in the day 3 fear context had significantly more correlated pairs than the highest degree node in the day 3 neutral context, significant

across $n = 5$ mice (58.8 versus 33.2 pairs, $P < 0.01$, paired t -test). **d**, Temporal relationship of HC neuron activity onset (set to time 0) compared with onset activity of correlated pairs (binned into 333 ms preceding or succeeding HC activity); $n = 48$ HC neurons. HCs were more likely to lead than lag their correlated pairs ($58.5 \pm 20\%$ leading versus $24.4 \pm 10\%$ lagging; $P < 0.01$, unpaired t -test). **e**, Significant synchronous activity (defined in Methods) quantified across five mice: number of synchronous events in the fear context was significantly greater than in the neutral context (5.8 ± 2.9 events in fear context versus 1.2 ± 1.1 in neutral context; $P < 0.01$, paired t -test). * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.



Extended Data Figure 9 | PCA of population trajectories in fear versus neutral contexts. **a**, PCA of $\Delta F/F$ traces of all active cells for mouse 2, performed separately for fear and neutral contexts. Population trajectories in the fear context take large, nearly orthogonal, deviations from the centre, while neutral context trajectories remain close to the origin. **b**, Three-dimensional reconstruction of the neuronal population showing that neurons participating in each synchronous event (red cells ($t = 39$), green circles ($t = 92$), blue circles

($t = 200$)) are largely non-overlapping and anatomically homogeneously distributed throughout the volume. There are, however, a small fraction of neurons that participate in all three events (black circles). **c**, **d**, Data are shown for another representative mouse. Similar results were seen in all other mice. **e**, The $\Delta F/F$ traces for a randomly selected set of 30 neurons participating in each of three events are shown, with the greatest amount of overlap seen between $t = 65$ and $t = 210$. Scale bars: 400% $\Delta F/F$, 20 s.



Extended Data Figure 10 | bReaChES: engineering a red-shifted opsin for robust projection targeting. **a**, Schematics of ReaChR³¹ and bReaChES.

ReaChR is a hybrid of segments from channelrhopsin-1 (blue, amino acids 1–95), *Volvox* channelrhodopsin-1 (red, amino acids 96–246, 279–350) and *Volvox* channelrhodopsin-2 (green, amino acids 247–278). The VChR1 segment contains the point mutation Leu171Ile. ReaChR was modified here for enhanced expression and membrane trafficking as well as accelerated channel kinetics, resulting in bReaChES, as follows. The first 51 amino-terminal residues were replaced by the first 11 amino-terminal residues from channelrhodopsin-2 (yellow, amino acids 1–11) and the last 5 carboxy-terminal residues were removed. Mutation of Glu 123 to Ser increases speed of channel closure. **b**, Spectra of C1V1_{TT}, bReaChES and ChR2 measured between 400 and 650 nm at 0.65 mW mm⁻² in cultured neurons from rat hippocampus (*n* = 6 each). **c**, Stationary photocurrents at 575 nm (C1V1_{TT} 630 ± 109 pA (s.e.m. throughout figure), ReaChR 963 ± 113 pA, bReaChES 1,365 ± 128 pA) and 632 nm (C1V1_{TT} 315 ± 111 pA, ReaChR 1,003 ± 95 pA, bReaChES 841 ± 102 pA). Current amplitudes were measured at -80 mV

and 5 mW mm⁻² light intensity, respectively. **d**, Speed of channel closure: τ value of mono-exponential off-kinetics (C1V1_{TT} 79 ± 3.7 ms, *n* = 26; ReaChR 682 ± 86 ms, *n* = 6; bReaChES 49 ± 4.4 ms, *n* = 25; *P* < 0.0005). **e**, **f**, Representative current-clamp traces of ReaChR- or bReaChES-expressing cultured neurons excited with 633 nm light (5 ms, 5 mW mm⁻²). ReaChR kinetics were slow enough that reliable action potential generation was only possible at very low frequencies (**e**), while the accelerated channel closure of bReaChES allowed reliable spike generation up to 20 Hz (**f**). **g**, **h**, Representative voltage-clamp (**g**) and current-clamp (**h**) traces of postsynaptic cells responding to light stimulation (orange) of bReaChES-expressing presynaptic terminals. Pulse length: 5 ms. **i**, **j**, Stationary photocurrents (**i**) and light-evoked spike probability (**j**) in opsin-expressing medial PFC (mPFC) cells in acute slice (C1V1_{TT}: *n* = 11, bReaChES: *n* = 10). **k**, **l**, Light-evoked EPSC amplitude (**k**) and spike probability (**l**) in postsynaptic cells (C1V1_{TT}: *n* = 10, bReaChES: *n* = 18). Light wavelength 575 nm (25 nm bandwidth) and power density 5 mW mm⁻². **P* < 0.05, ***P* < 0.01, ****P* < 0.001.

Cleavage of GSDMD by inflammatory caspases determines pyroptotic cell death

Jianjin Shi^{1,2*}, Yue Zhao^{2*}, Kun Wang², Xuyan Shi², Yue Wang², Huanwei Huang², Yinghua Zhuang², Tao Cai², Fengchao Wang² & Feng Shao^{2,3,4}

Inflammatory caspases (caspase-1, -4, -5 and -11) are critical for innate defences. Caspase-1 is activated by ligands of various canonical inflammasomes, and caspase-4, -5 and -11 directly recognize bacterial lipopolysaccharide, both of which trigger pyroptosis. Despite the crucial role in immunity and endotoxic shock, the mechanism for pyroptosis induction by inflammatory caspases is unknown. Here we identify gasdermin D (*Gsdmd*) by genome-wide clustered regularly interspaced palindromic repeat (CRISPR)-Cas9 nuclease screens of caspase-11- and caspase-1-mediated pyroptosis in mouse bone marrow macrophages. *GSDMD*-deficient cells resisted the induction of pyroptosis by cytosolic lipopolysaccharide and known canonical inflammasome ligands. Interleukin-1 β release was also diminished in *Gsdmd*^{-/-} cells, despite intact processing by caspase-1. Caspase-1 and caspase-4/5/11 specifically cleaved the linker between the amino-terminal gasdermin-N and carboxy-terminal gasdermin-C domains in *GSDMD*, which was required and sufficient for pyroptosis. The cleavage released the intramolecular inhibition on the gasdermin-N domain that showed intrinsic pyroptosis-inducing activity. Other gasdermin family members were not cleaved by inflammatory caspases but shared the autoinhibition; gain-of-function mutations in *Gsdma3* that cause alopecia and skin defects disrupted the autoinhibition, allowing its gasdermin-N domain to trigger pyroptosis. These findings offer insight into inflammasome-mediated immunity/diseases and also change our understanding of pyroptosis and programmed necrosis.

Inflammatory caspases (caspase-1, murine caspase-11 and human caspase-4/5) are crucial for innate immune defences. Aberrant or excessive activation of caspase-1 causes or is associated with many autoinflammatory, autoimmune and even metabolic diseases^{1,2}. Caspase-1 is activated by the canonical inflammasomes, in which a central scaffold, such as NLRP3, NLRP1, NAIP-NLRC4, AIM2 and Pyrin, detects its cognate ligand¹. The NAIPs directly recognize flagellin, as well as the rod and needle components of bacterial type III secretion system (T3SS)^{3–5}. The newly identified Pyrin inflammasome indirectly senses bacterial modifications and inactivation of host Rho GTPases⁶. Caspase-11 detects cytosolic bacterial lipopolysaccharide (LPS), playing a critical role in endotoxic shock^{7,8}. Caspase-11 and its human counterparts caspase-4/5 are activated by direct LPS binding^{9,10}.

Caspase-1 activation mainly occurs in macrophage/dendritic cells. Active caspase-1 processes and maturates interleukin (IL)-1 β /18 and also triggers a form of programmed necrosis known as pyroptosis. Pyroptosis is the dominant response upon caspase-4/5/11 activation, occurring in both macrophage and non-macrophage cells^{9,11}. Pyroptosis features pore formation in the plasma membrane, cell swelling and rupture of the membrane, causing massive leakage of cytosolic contents. Accumulating evidences suggest a critical role of pyroptosis in immunity and disease¹². Pyroptosis of infected cells releases intracellular bacteria for neutrophil-mediated killing¹³. Inflammasome-mediated, but IL-1 β /18-independent, clearance of infection has been noted with several intracellular bacteria^{12–15}. Pyroptosis induced by the NLRP1B inflammasome or a gain-of-function mutation in *Nlrp1a* is the primary cause of anthrax-lethal-toxin-induced lung injury¹⁶ and haematopoietic progenitor cell

depletion in mice¹⁷, respectively. Caspase-1-mediated pyroptosis in HIV infection accounts for CD4⁺ T-cell depletion¹⁸, a critical event in HIV pathogenesis. Pyroptosis rather than cytokine secretion is probably the key determinant of mouse lethality caused by systemic activation of the NAIP-NLRC4 inflammasome or caspase-11 activation by excessive LPS^{7,8,19}. Despite these important functions, the mechanism underlying how inflammatory caspases trigger pyroptosis is unknown.

GSDMD is required for LPS-induced pyroptosis

We previously established an LPS electroporation protocol that can induce pyroptosis in more than 90% of LPS-stimulated cells⁹. This robust assay allowed for unbiased genome-wide genetic screen using the CRISPR-Cas9 technology to identify new components in LPS/caspase-4/11-induced pyroptosis. The screen was performed in *Tlr4*^{-/-} immortalized bone-marrow-derived macrophages (iBMDMs) that responded normally to LPS electroporation (Extended Data Fig. 1a). As expected, three out of the four *Casp11*-targeting guide RNAs (gRNAs) were recovered within the top 100 hits. Notably, there was only one other gene gasdermin D (*Gsdmd*) that also had multiple gRNA hits; four out of the five *Gsdmd*-targeting gRNAs appeared in the top 30 hits, with two among the top 10 (Fig. 1a). *GSDMD* is conserved in human and mouse with ~72% sequence similarity (Extended Data Fig. 2a). Small interference RNA (siRNA) knockdown in HeLa cells excluded other selected top hits but confirmed the requirement of *GSDMD* (Extended Data Fig. 1b–d, also see Methods). siRNA knockdown of *Gsdmd* in mouse iBMDMs also inhibited LPS-induced pyroptosis; the extent of inhibition by three different siRNAs correlated with their knockdown efficiency (Extended Data Fig. 1e, f).

¹Peking University-Tsinghua University-National Institute of Biological Sciences Joint Graduate Program, School of Life Sciences, Tsinghua University, 100084, China. ²National Institute of Biological Sciences, Beijing 102206, China. ³National Laboratory of Biomacromolecules, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China. ⁴National Institute of Biological Sciences, Beijing, Collaborative Innovation Center for Cancer Medicine, Beijing 102206, China.

*These authors contributed equally to this work.

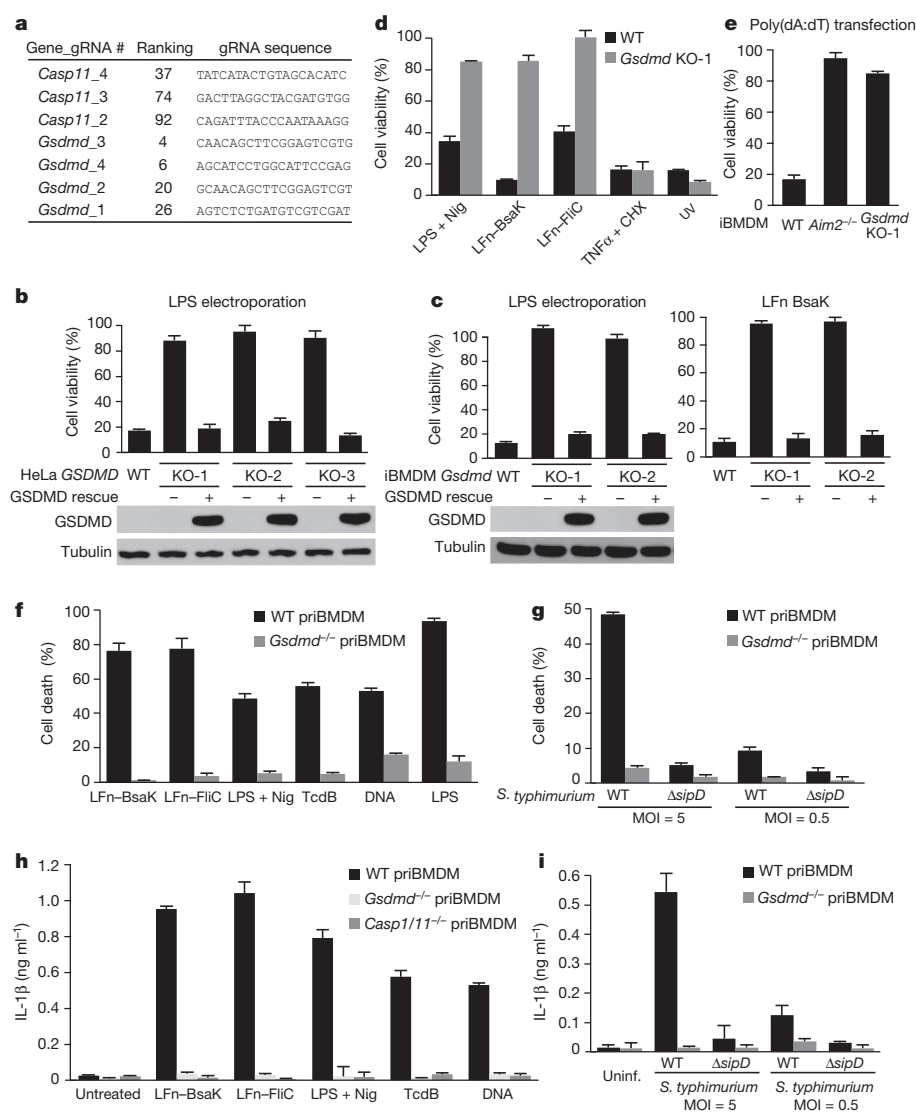


Figure 1 | Genetic screens identify that GSDMD is required for inflammatory-caspase-mediated pyroptosis and its role in caspase-1-induced IL-1 β release. **a**, gRNA hits from a genome-wide CRISPR-Cas9 screen of LPS-induced pyroptosis in mouse iBMDMs. Shown are genes with multiple gRNA hits among the top 100 hits. Fold-enrichment-based ranking and the target sequence of each gRNA are listed. **b**, **c**, Effects of GSDMD knockout on LPS electroporation and LFn-BsaK-induced pyroptosis. Three *GSDMD*^{-/-} HeLa clones (KO-1, -2 and -3) and two *Gsdmd*^{-/-} iBMDM clones generated by CRISPR-Cas9-mediated targeting were analysed. For complementation, Flag-tagged human GSDMD was stably expressed in the knockout cells and its expression is shown by anti-Flag immunoblotting; tubulin expression is the loading control. WT, wild type. **d**, **e**, Effects of *Gsdmd* knockout on canonical inflammasome-mediated pyroptosis or on apoptosis. LFn-BsaK, LFn-FliC (flagellin), LPS + Nig. (nigericin), poly(dA:dT) transfection, TNF α + CHX and ultraviolet (UV) radiation were employed to activate the NAIP2-NLRC4, the NAIP5-NLRC4, the NLRP3, the AIM2 inflammasome, the extrinsic and the intrinsic apoptosis pathways, respectively. **f**–**i**, Assays for inflammasome-mediated pyroptosis and IL-1 β release. Primary BMDMs (priBMDMs) from wild-type or *Gsdmd*^{-/-} mice were stimulated with indicated canonical inflammasome stimuli or LPS electroporation (**f**, **h**), or infected with *S. typhimurium* (wild-type or the T3SS-deficient Δ sidP mutant) at the indicated multiplicity of infection (MOI) (**g**, **i**). TcdB, *C. difficile* toxin B; Uninf., uninfected cells. ATP-based cell viability (**b**–**e**), lactate dehydrogenase (LDH)-release-based cell death (**f**, **g**), and ELISA assay of IL-1 β release (**h**, **i**) are expressed as mean values \pm s.d. from three technical replicates. Data shown are representative of at least three (**b**–**e**) or two (**f**–**i**) independent experiments.

To further validate the role of GSDMD, *GSDMD*^{-/-} HeLa cells and *Gsdmd*^{-/-} iBMDM cells were generated by CRISPR-Cas9-mediated targeting (Extended Data Fig. 3a). In both cells, the absence of GSDMD completely blocked LPS electroporation-triggered pyroptosis, which was confirmed in multiple knockout clones (Extended Data Fig. 3b, c). Re-expression of human GSDMD in three *GSDMD*^{-/-} HeLa cell clones all restored LPS-induced pyroptosis (Fig. 1b). Mouse GSDMD showed the same rescue effect (Extended Data Fig. 3d). Similarly, *Gsdmd*^{-/-} iBMDM cells regained the sensitivity to LPS electroporation when complemented with either human or mouse GSDMD (Fig. 1c and Extended Data Fig. 3e).

GSDMD is required for caspase-1-mediated pyroptosis

We also performed a parallel CRISPR-Cas9 screen on caspase-1-mediated pyroptosis in iBMDM cells using the anthrax lethal factor N-terminal-domain-fused T3SS rod protein of *Burkholderia thailandensis* (LFn-BsaK), the most potent agonist of the NAIP-NLRC4 inflammasome^{4,20}. Notably, the top hits recovered from the screen were dominated by gRNAs targeting all known components in the pathway, including *Naip2*, *Nlr4* and *Casp1*, each of which was hit by multiple gRNAs (Extended Data Fig. 4a). The screen also hit *Gsdmd* which, along with seven other genes, had two independent gRNAs (Extended Data Fig. 4a). The requirement of *Gsdmd*, but not the seven other genes, for LFn-BsaK-induced pyroptosis was confirmed by siRNA knockdown assays (Extended Data Fig. 1e, f and 4b).

Furthermore, *Gsdmd*^{-/-} iBMDM clones completely resisted LFn-BsaK-induced pyroptosis (Extended Data Fig. 3f), which was fully restored by exogenous expression of GSDMD (Fig. 1c). *Gsdmd*^{-/-} iBMDMs were also unresponsive to stimulation by LFn-flagellin (FliC), an agonist of the NAIP5-NLRC4 inflammasome (Fig. 1d). Caspase-7 was previously shown to be cleaved by caspase-1 upon inflammasome activation by flagellin in *Legionella pneumophila*-infected mouse macrophages²¹. We confirmed the cleavage of caspase-7/3 in LFn-BsaK-stimulated *Gsdmd*^{-/-} iBMDMs (Extended Data Fig. 4c). However, evident cleavage could be observed 6 h after stimulation, a time point when *Gsdmd*-proficient cells had already died of pyroptosis (therefore little caspase-3/7 cleavage occurred). This suggests that caspase-3/7 cleavage plays no role in caspase-1-mediated pyroptosis, but may induce apoptosis or have other functions in defence against *L. pneumophila*²¹. Thus, GSDMD plays a critical role in caspase-1-mediated pyroptosis downstream of the NAIP-NLRC4 inflammasome.

GSDMD is only required for pyroptotic cell death

We further examined the role of GSDMD in other canonical-inflammasome-triggered pyroptosis. *Gsdmd*^{-/-} completely blocked iBMDM cell pyroptosis upon activation of the NLRP3 and AIM2 inflammasomes by LPS plus nigericin and poly(dA:dT), respectively (Fig. 1d, e). In contrast, *Gsdmd*^{-/-} did not affect extrinsic and intrinsic apoptosis triggered by TNF α plus cycloheximide (TNF α + CHX) and

ultraviolet irradiation, respectively (Fig. 1d). siRNA knockdown of *GSDMD* in HT-29 cells, though efficiently blocking LPS-induced pyroptosis, had no effect on TSZ (TNF α plus SMAC mimetic and the caspase inhibitor zVAD)-induced necroptosis, another form of programmed necrosis mediated by the RIPK1–RIPK3–MLKL axis²² (Extended Data Fig. 5a–c).

We generated *GSDMD*-deficient mice by CRISPR–Cas9-mediated targeting. Primary BMDMs from two independent homozygous mutant mice (F1-1 and F1-2) (Extended Data Fig. 6a) were analysed for inflammasome activation, and similar results were obtained. In the canonical inflammasome pathway, pyroptosis mediated by the NAIP2/5–NLRC4, NLRP3 and AIM2 inflammasomes were all blocked by *GSDMD* deficiency (Fig. 1f). *Gsdmd*^{−/−} also abolished pyroptosis upon *Clostridium difficile* toxin B activation of the Pyrin inflammasome⁶. Similarly, *Gsdmd*^{−/−} BMDMs resisted LPS-induced pyroptosis (Fig. 1f). When infected with *Salmonella typhimurium*, wild-type BMDMs showed extensive pyroptosis due to T3SS-dependent activation of NAIP–NLRC4 inflammasome, but little pyroptosis occurred in BMDMs from *Gsdmd*^{−/−} mice (Fig. 1g). Defective pyroptosis in *Gsdmd*^{−/−} BMDMs was not due to the absence of caspase-1/11 expression (Extended Data Fig. 6b). Consistent with that in iBMDMs, primary *Gsdmd*^{−/−} BMDMs remained sensitive to necroptosis induction by TSZ or LPS plus zVAD (Extended Data Fig. 5d, e). These data highlight a specific role of *GSDMD* in caspase-4/11 and caspase-1-mediated pyroptosis downstream of multiple inflammasomes.

GSDMD controls IL-1 β release but not maturation

IL-1 β maturation and secretion is the other major response of canonical inflammasome activation. We observed that primary *Gsdmd*^{−/−} BMDMs, in contrast to wild-type BMDMs, secreted little IL-1 β into the supernatant upon activation of the NAIP2/5–NLRC4, NLRP3, Pyrin and AIM2 inflammasomes (Fig. 1h). *S. typhimurium*-infection-induced IL-1 β release was also inhibited in *Gsdmd*^{−/−} BMDMs (Fig. 1i). Notably, caspase-1 autoprocessing occurred normally in *Gsdmd*^{−/−} BMDMs upon activation of the above inflammasomes while extracellular secretion of both pro-caspase-1 and the mature caspase-1 was severely inhibited (Extended Data Fig. 6c). Intact caspase-1 autoprocessing and IL-1 β maturation were also observed in the cytosol of *Gsdmd*^{−/−} iBMDMs upon canonical inflammasomes activation (Fig. 2a, b and Extended Data Fig. 7a). Thus, *GSDMD*-mediated pyroptosis plays an important role in mature IL-1 β release without affecting its maturation.

GSDMD is specifically cleaved by caspase-1/4/5/11

The above analyses predict that *GSDMD* functions downstream of the inflammatory caspases. We observed that *GSDMD* was cleaved in LPS and LFn–BsaK-stimulated HeLa and iBMDM cells, respectively (Fig. 2c). A cleavage product (~38 kDa) corresponding to the N-terminal half of 2 \times Flag–HA–*GSDMD* was constantly identified. LPS and LFn–BsaK-induced *GSDMD* cleavage was diminished by a pan-caspase inhibitor zVAD that also blocked cell death (Extended Data Fig. 7b). Co-expression of *GSDMD* with caspase-1, 4, 5 or 11 but not apoptotic caspases (caspase-2, 8 and 9) in 293T cells induced the same cleavage of *GSDMD* (Extended Data Fig. 7c). No *GSDMD* cleavage occurred in TNF α + CHX and TSZ-stimulated cells (Extended Data Fig. 5f), consistent with that *GSDMD* was not required for apoptosis and necroptosis (Fig. 1d and Extended Data Fig. 5c).

We then tested whether *GSDMD* is a direct substrate of inflammatory caspases. The active tetramer forms of caspase-1, 4 and 11, but not caspase-2, 8 and 9, were found capable of cleaving purified recombinant *GSDMD* (no tag, ~53 kDa) (Fig. 2d). The larger-size cleavage product migrated at ~31 kDa, which plus the 7-kDa 2 \times Flag–HA tag equals the 38-kDa N-terminal cleavage product observed in cells (Fig. 2c). The other cleavage product (~22 kDa) matched the size of the remaining C-terminal fragment. Consistent with that caspase-4/11

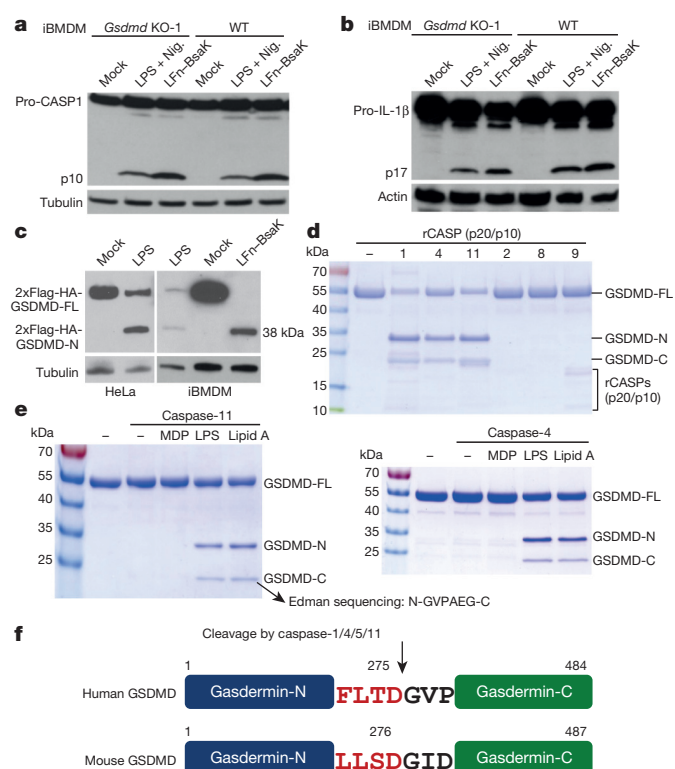


Figure 2 | GSDMD is a substrate specifically for inflammatory caspases.

a, b, Effects of *Gsdmd* knockout on intracellular processing of caspase-1 and IL-1 β (stably expressed). p10, mature caspase-1; Pro-CASP1, caspase-1 precursor; Pro-IL-1 β , IL-1 β precursor; p17, mature IL-1 β . **c**, A specific cleavage in *GSDMD* induced by LPS electroporation and LFn–BsaK. *GSDMD*-FL, full-length *GSDMD*. **d**, *In vitro* cleavage of *GSDMD* by recombinant caspases. Purified *GSDMD* was incubated with active p20/p10 tetramers of indicated caspases (rCASP). **e**, *In vitro* cleavage of *GSDMD* by LPS-activated caspase-11/4. Purified *GSDMD* was incubated with insect-cell-purified caspase-11/4 pre-incubated with LPS or indicated ligands. The N-terminal sequence of *GSDMD*-C determined by Edman sequencing is shown. **f**, Cartoon diagram of *GSDMD* structure and the cleavage by inflammatory caspases. LFn–BsaK and LPS + Nig. were used to activate the NAIP2–NLRC4 and NLRP3 inflammasome, respectively (**a–c**). 2 \times Flag–HA–*GSDMD* was stably expressed in HeLa and iBMDM cells (**c**). Reaction samples were analysed by SDS–PAGE and Coomassie blue staining (**d, e**). Cell lysates were subjected to anti-caspase-1 (**a**), anti-IL-1 β (**b**), anti-Flag (**c**), anti-tubulin (**a, c**) and anti-actin (**b**) immunoblotting. All data shown are representative of three independent experiments.

are activated by direct LPS binding⁹, insect-cell-purified caspase-11 could process recombinant *GSDMD* into the 31 kDa and 22 kDa fragments only upon prior incubation with LPS or Lipid A but not muramyl dipeptide (MDP) (Fig. 2e). Identical results were observed with insect-cell-purified caspase-4. Edman sequencing of the C-terminal fragment revealed an N-terminal sequence of 276GVPAEG₂₈₁ (Fig. 2e). Preceding 276GVPAEG₂₈₁ is 272FLTD₂₇₅ in human *GSDMD* and 273LLSD₂₇₆ in mouse *GSDMD* (Fig. 2f), which fit well with the substrate recognition motif of inflammatory caspases²³. 272FLTD₂₇₅ is the only possible inflammatory-caspase cleavage site conserved in human and mouse *GSDMD* (Extended Data Fig. 2a). Thus, inflammatory caspases specifically cleave *GSDMD* after the 272FLTD₂₇₅ (or 273LLSD₂₇₆) sequence (Fig. 2f).

Cleavage of *GSDMD* is required for pyroptosis

We generated a mutant *GSDMD*, in which Asp275 (Asp276 in mouse *GSDMD*) was replaced with an alanine. As expected, the D/A mutants of *GSDMD* resisted cleavage by caspase-1 and caspase-4/5/11 over-expressed in 293T cells (Extended Data Fig. 8a, b). Physiologically

activated caspase-1 by LFn-BsaK also failed to cleave GSDMD D275A complemented into the *Gsdmd*^{-/-} iBMDMs (Fig. 3a); importantly, the mutant was completely inactive in restoring LFn-BsaK-induced pyroptosis (Fig. 3b). GSDMD D275A also resisted cleavage by caspase-11 and caspase-4 in LPS-stimulated iBMDM and HeLa cells, respectively (Fig. 3a, c), and was unable to mediate LPS-induced pyroptosis (Fig. 3b, d).

In bacterial infection assays, complementation of the *Gsdmd*^{-/-} iBMDMs with wild-type GSDMD restored *S. typhimurium*-induced pyroptosis but the D275A mutant showed no activity (Fig. 3e). While wild-type GSDMD was cleaved in the T3SS-dependent manner, the D275A mutant protein remained intact (Extended Data Fig. 8c). The same results were observed in *B. thailandensis* and enteropathogenic *Escherichia coli* infections (Fig. 3f and Extended Data Fig. 8d). In HeLa cell infection, *S. typhimurium* Δ sifA triggered

caspase-4-dependent pyroptosis^{9,14}. This cell death was diminished in *GSDMD*^{-/-} cells (Fig. 3g). Notably, expression of wild-type GSDMD but not the D275A mutant in *GSDMD*^{-/-} cells restored *S. typhimurium* Δ sifA-triggered pyroptosis (Fig. 3g).

Cleavage of GSDMD is sufficient to drive pyroptosis

To investigate whether cleavage at Asp275 in GSDMD is sufficient to initiate pyroptosis, a PreScission protease (PPase) recognition sequence (EVLFGNP) was inserted between Leu273 and Thr274 in the 272FLTD₂₇₅ motif in GSDMD. Expression of this engineered GSDMD(EVLFGNP) mutant or wild-type GSDMD in 293T cells by itself triggered no pyroptosis. However, upon cytosolic delivery of purified PPase, cells expressing GSDMD(EVLFGNP) but not wild-type GSDMD died extensively (Fig. 4a). Consistently, GSDMD(EVLFGNP) but not the wild-type protein was cleaved into the expected size by the PPase (Fig. 4b). The dying cells developed typical pyroptosis morphology with cell swelling and membrane rupture (Fig. 4a). We also

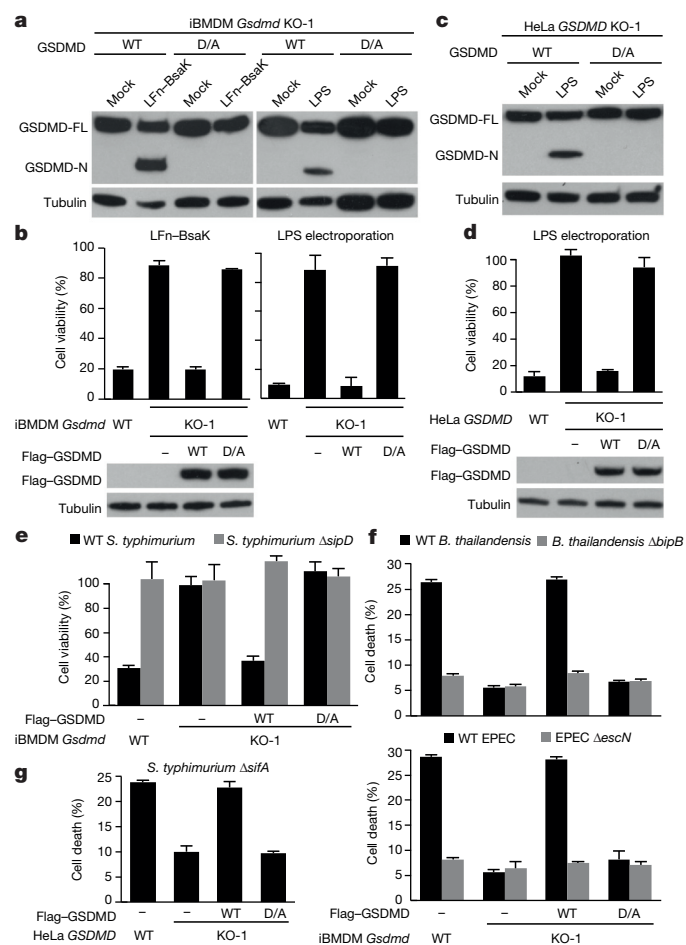


Figure 3 | Cleavage of GSDMD by inflammatory caspases is required for pyroptosis. **a–d**, Assays of the cleavage-resistant mutant of GSDMD in caspase-1 and caspase-4/11-mediated pyroptosis induced by LFn-BsaK and LPS electroporation, respectively. **e–g**, Assays of the cleavage-resistant mutant of GSDMD in bacterial-infection-induced pyroptosis. iBMDM cells were infected with wild-type *S. typhimurium* (**e**), *B. thailandensis* or enteropathogenic *E. coli* (EPEC) (**f**) or their T3SS-deficient mutants (Δ sifD, Δ bipB and Δ escN, respectively) as controls. HeLa cells were infected with *S. typhimurium* Δ sifA (**g**). The wild-type or the D275A mutants (D/A) of 2×Flag-HA-GSDMD (**a, c**) and Flag-GSDMD (**b, d–g**) were stably expressed in the knockout cells. Cell lysates were analysed by anti-Flag and anti-tubulin immunoblotting (**a–d**). Lysates of non-stimulated cells were analysed in **b** and **d**. ATP-based cell viability (**b, d, e**) and LDH-release-based cell death (**f, g**) are expressed as mean values \pm s.d. from three technical replicates. All data shown are representative of three independent experiments.

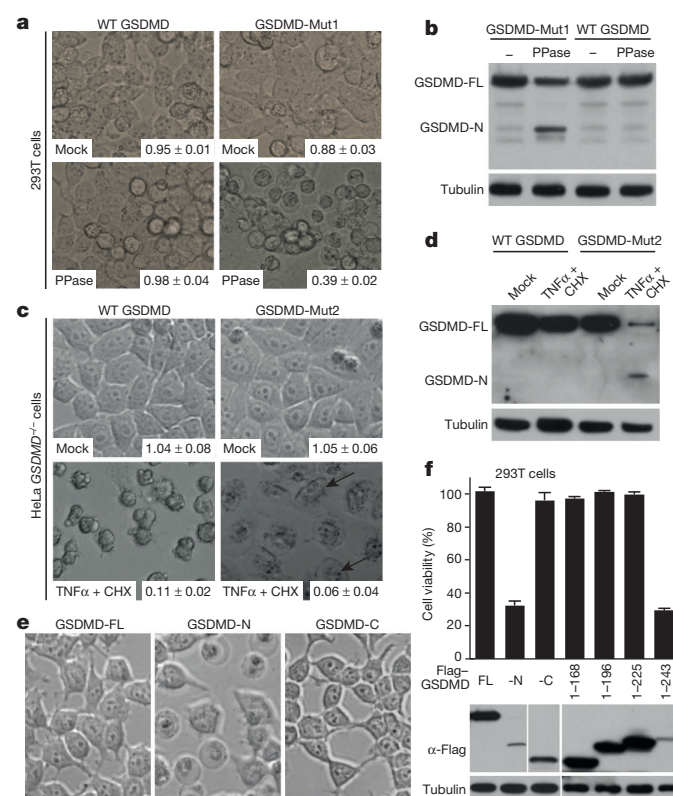


Figure 4 | Interdomain cleavage of GSDMD is sufficient to trigger pyroptosis owing to the intrinsic pyroptosis-inducing activity in its N-terminal domain. **a, b**, Pyroptosis induced by interdomain cleavage of an engineered GSDMD by PPase. 293T cells were transfected with Flag-tagged wild-type GSDMD or GSDMD-Mut1 (insertion of the PPase recognition sequence EVLFQGP into the FLTD site). Purified PPase was electroporated into the cells. **c, d**, Switch of TNF α -induced apoptosis to pyroptosis by the caspase-3-sensitive GSDMD mutant. Flag-tagged wild-type GSDMD or GSDMD-Mut2 (replacement of the FLTD motif with the caspase-3-cleavage site DEVG) was stably expressed in HeLa *GSDMD*^{-/-} cells. Cells were treated with TNF α + CHX for 8 h. **e, f**, The pyroptosis-inducing activity of GSDMD N-terminal domain. Indicated GSDMD fragments were transiently expressed in 293T cells. GSDMD-FL, full-length GSDMD; GSDMD-N and GSDMD-C, the N- and C-terminal product of GSDMD generated by inflammatory caspase cleavage, respectively. Phase-contrast images of cell morphology are shown in **a, c** and **e**. ATP-based cell viability is expressed as mean values \pm s.d. from three technical replicates (**a, c, f**) and are marked on the images in decimal form (**a, c**). Cell lysates were analysed by anti-Flag and anti-tubulin immunoblotting (**b, d, f**). All data shown are representative of three independent experiments.

engineered the FLTD site into DEVD, the cleavage site of caspase-3/7, and the resulting GSDMD(DEVD) mutant was stably expressed in *GSDMD*^{-/-} HeLa cells. When stimulated with TNF α + CHX, both wild-type GSDMD and GSDMD(DEVD) mutant-expressing cells showed massive cell death (Fig. 4c). As expected, cells expressing wild-type GSDMD underwent apoptosis with no GSDMD cleavage. In contrast, GSDMD(DEVD) mutant-expressing cells showed typical pyroptotic morphology, in which the mutant GSDMD was proteolysed to the expected size by activated endogenous caspase-3/7 (Fig. 4c, d). The switch of TNF α -induced apoptosis to pyroptosis by the caspase-3/7-sensitive GSDMD(DEVD) can be observed in Supplementary Videos 1 and 2. These results establish that proteolytic cleavage at Asp275 in GSDMD is sufficient to instruct mammalian cells to undergo pyroptosis.

Pyroptosis-inducing activity of GSDMD-N domain

The function of GSDMD is completely unknown. *Gsdmd*^{-/-} mice have no obvious phenotypes with normal epithelial differentiation in the intestinal tract despite high expression in gastrointestinal tracts²⁴. To understand the functional mechanism of GSDMD cleavage in pyroptosis, the N-terminal and C-terminal cleavage fragments (GSDMD-N and GSDMD-C, respectively) were individually expressed in 293T cells. In contrast to GSDMD-C and full-length GSDMD, GSDMD-N could cause extensive cell death with apparent pyroptosis morphology (Fig. 4e, f). GSDMD-N was expressed at a much lower level than GSDMD-C and full-length GSDMD, a phenomenon commonly seen in proteins with cytotoxicity. Progressive truncations identified residues 1–243 of GSDMD as the minimal fragment capable of triggering pyroptosis (Fig. 4f).

Autoinhibition of GSDMD and the gasdermin family

Above analyses suggest that GSDMD-C may interact with GSDMD-N and this autoinhibition is released upon interdomain cleavage by inflammatory caspases. Supporting this idea, GSDMD-C was efficiently co-immunoprecipitated by GSDMD-N in transfected 293T cells (Fig. 5a). Furthermore, overexpression of GSDMD-C could block LPS-induced pyroptosis in HeLa cells due to *trans*-inhibition of endogenous GSDMD-N generated from caspase-4 cleavage (Fig. 5b).

GSDMD belongs to the gasdermin family that also contains GSDMA, GSDMB, GSDMC, DFNA5 and DFNB59 (Extended Data

Fig. 2b)^{25,26}. The gasdermin family shares ~ 45% overall sequence homology. Structures of the gasdermins can be divided into two domains, the gasdermin-N and -C domains, corresponding to the two cleavage fragments of GSDMD by inflammatory caspases (Fig. 2f). Mice lack *Gsdmb*, but encode three GSDMA (GSDMA1–A3) and four GSDMC (GSDMC1–C4) proteins (Extended Data Fig. 2b). Other gasdermins do not share the FLTD motif in GSDMD (Extended Data Fig. 2a). Consistently, GSDMA, GSDMB and GSDMC resisted caspase-1/11 cleavage in 293T cells (Extended Data Fig. 9a). The gasdermin-N domain of GSDMA, GSDMB and GSDMC all could bind to its respective gasdermin-C domain (Fig. 5c), suggesting a similar autoinhibition that is released by mechanisms other than caspase cleavage.

Pyroptosis-inducing activity of GSDMA3 in disease

The function of the gasdermin family is poorly characterized^{25,26}. GSDMA3 is the most studied gasdermin; mice with spontaneous or chemically induced mutations in *Gsdma3* exhibit hyperkeratosis and hair-loss phenotypes in the skin^{25–27}. The mechanism for this alopecia pathology is unknown, but severe chronic inflammation has been observed in the skin of *Gsdma3*-mutant mice^{28,29}. Like GSDMD, GSDMA3 exhibited the intramolecular interaction between its gasdermin-N and -C domains (Fig. 5c). Overexpression of the gasdermin-N domain of GSDMA3 (residues 1–284), but not full-length GSDMA3, caused extensive pyroptosis in 293T cells (Fig. 5d); GSDMA3 was not cleaved by inflammatory caspases but artificial interdomain cleavage of an engineered GSDMA3 by PPase could cause 293T cell pyroptosis (Extended Data Fig. 9b). GSDMA3 also remained intact in TNF α -induced apoptosis and necroptosis (Extended Data Fig. 9c). These data suggest that GSDMA3 is an autoinhibited pyroptosis-inducing factor functioning in an unknown biological process. The autoinhibition is partially supported by a recent intriguing study that proposes a function of GSDMA3 in autophagy³⁰.

Gsdma3^{-/-} mice show no visible skin defects³⁰. In fact, the hyperkeratosis and hair-loss phenotype in *Gsdma3*-mutant mice is dominant and the mutations are gain-of-function²⁵. There are nine reported *Gsdma3* mutant alleles, resulting in 259RDW (insertion after residue 259 with mistranslated RDW sequence), T278P, L343P, Y344C, Y344H, A348T, I359N, premature stop at 366, and duplica-

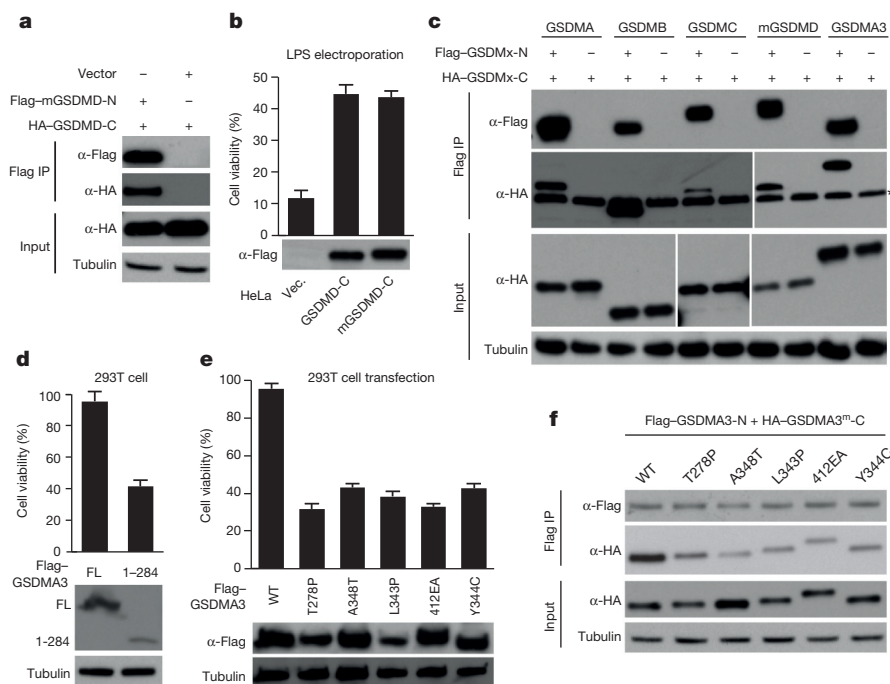


Figure 5 | Autoinhibition of the pyroptosis-inducing activity of the gasdermin family.

a, c, Co-immunoprecipitation assay of the interaction between the gasdermin-N and -C domains in GSDMD and other gasdermin family members. * indicates antibody light chain. mGSDMD, mouse GSDMD. **b**, Effects of GSDMD-C domain overexpression on LPS electroporation-induced pyroptosis. Vec., vector. **d, e**, The pyroptosis-inducing activity of GSDMA3-N domain and its alopecia-causing mutants. **f**, Effects of the alopecia-causing mutations on the co-immunoprecipitation interaction between the N and C domains of GSDMA3. GSDMA3^{mut}, point mutants of GSDMA3 analysed here. Indicated gasdermin constructs were expressed in 293T cells. Anti-Flag immunoprecipitation (**a, c, f**) was performed. The immunoprecipitates (Flag IP) or the total lysates were analysed by immunoblotting with antibodies against indicated epitopes or tubulin. ATP-based cell viability is expressed as mean values \pm s.d. from three technical replicates (**b, d, e**). All data shown are representative of three independent experiments.

tion of E411A412 (412EA), in the gasdermin-C domain^{25–27}. Five of the mutations (T278P, L343P, Y344C, A348T and 412EA) were analysed and found capable of instructing 293T cells to undergo pyroptosis (Fig. 5e). These mutations all disrupted or attenuated the co-immunoprecipitation between the gasdermin-N and -C domains of GSDMA3 (Fig. 5f). Thus, the alopecia-causing GSDMA3 mutants lose the autoinhibition and become constitutively active in triggering pyroptosis. Taken all together, our analyses further suggest that gasdermins are all pyroptosis-inducing factors that are activated by release of the autoinhibition on their gasdermin-N domains.

Discussion

We identify GSDMD as a generic substrate for caspase-1 and caspase-4/5/11. Cleavage of GSDMD by the inflammatory caspases critically determines pyroptosis by releasing the cleaved gasdermin-N domain that bears intrinsic pyroptosis-inducing activity. Further understanding the mechanism of action of GSDMD may provide a new avenue for therapeutic intervention of inflammatory-caspase-associated autoinflammatory conditions and septic shock. We also show that other gasdermins have a similar pyroptosis function; importantly, this observation redefines the concept of pyroptosis as gasdermin-mediated programmed necrosis.

Caspases are classified into apoptotic and inflammatory caspases. Substrate targeting by these caspases is known to be promiscuous to some extent; overlapping substrate spectra for caspases of different natures have been noted in proteomic efforts that identified GSDMD as a substrate for both caspase-1 and apoptotic caspases^{31,32}. Our data suggest that interdomain cleavage of GSDMD by caspase-1/4/5/11 determines pyroptosis. Caspase-1 can also cleave caspase-3/7²¹, which may have a pyroptosis-independent function, particularly in the absence of GSDMD. Thus, the substrate cleaved by a caspase in a cellular context is the determinant for the nature of the cell death. Pyroptosis and necroptosis are two types of lytic and programmed necrosis. In necroptosis, a pseudokinase MLKL, upon phosphorylation by RIPK3, acts as the executioner to rupture the membrane²². It is unclear at this stage whether GSDMD plays a similar ‘executioner’ role in pyroptosis, but our data suggest that GSDMD, or more specifically its gasdermin-N domain, is sufficient to drive pyroptosis regardless of the cellular system. Given the possible presence of other GSDMD- and MLKL-like necrosis factors, our results suggest a new paradigm for understanding programmed necrosis.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 23 June; accepted 25 August 2015.

Published online 16 September 2015.

- Lamkanfi, M. & Dixit, V. M. Mechanisms and functions of inflammasomes. *Cell* **157**, 1013–1022 (2014).
- Henao-Mejia, J., Elinav, E., Thaïss, C. A. & Flavell, R. A. Inflammasomes and metabolic disease. *Annu. Rev. Physiol.* **76**, 57–78 (2014).
- Kofoed, E. M. & Vance, R. E. Innate immune recognition of bacterial ligands by NALPs determines inflammasome specificity. *Nature* **477**, 592–595 (2011).
- Zhao, Y. *et al.* The NLR4 inflammasome receptors for bacterial flagellin and type III secretion apparatus. *Nature* **477**, 596–600 (2011).
- Zhao, Y. & Shao, F. The NALP-NLR4 inflammasome in innate immune detection of bacterial flagellin and type III secretion apparatus. *Immunol. Rev.* **265**, 85–102 (2015).
- Xu, H. *et al.* Innate immune sensing of bacterial modifications of Rho GTPases by the P2Y₆ inflammasome. *Nature* **513**, 237–241 (2014).
- Kayagaki, N. *et al.* Noncanonical inflammasome activation by intracellular LPS independent of TLR4. *Science* **341**, 1246–1249 (2013).
- Hagar, J. A., Powell, D. A., Aachoui, Y., Ernst, R. K. & Miao, E. A. Cytoplasmic LPS activates caspase-11: implications in TLR4-independent endotoxemic shock. *Science* **341**, 1250–1253 (2013).

- Shi, J. *et al.* Inflammasome caspases are innate immune receptors for intracellular LPS. *Nature* **514**, 187–192 (2014).
- Yang, J., Zhao, Y. & Shao, F. Non-canonical activation of inflammatory caspases by cytosolic LPS in innate immunity. *Curr. Opin. Immunol.* **32**, 78–83 (2015).
- Kayagaki, N. *et al.* Non-canonical inflammasome activation targets caspase-11. *Nature* **479**, 117–121 (2011).
- Jorgensen, I. & Miao, E. A. Pyroptotic cell death defends against intracellular pathogens. *Immunol. Rev.* **265**, 130–142 (2015).
- Miao, E. A. *et al.* Caspase-1-induced pyroptosis is an innate immune effector mechanism against intracellular bacteria. *Nature Immunol.* **11**, 1136–1142 (2010).
- Aachoui, Y. *et al.* Caspase-11 protects against bacteria that escape the vacuole. *Science* **339**, 975–978 (2013).
- Sauer, J. D. *et al.* *Listeria monocytogenes* engineered to activate the Nlrp4 inflammasome are severely attenuated and are poor inducers of protective immunity. *Proc. Natl Acad. Sci. USA* **108**, 12419–12424 (2011).
- Kovarova, M. *et al.* NLRP1-dependent pyroptosis leads to acute lung injury and morbidity in mice. *J. Immunol.* **189**, 2006–2016 (2012).
- Masters, S. L. *et al.* NLRP1 inflammasome activation induces pyroptosis of hematopoietic progenitor cells. *Immunity* **37**, 1009–1023 (2012).
- Doitsh, G. *et al.* Cell death by pyroptosis drives CD4 T-cell depletion in HIV-1 infection. *Nature* **505**, 509–514 (2014).
- von Moltke, J. *et al.* Rapid induction of inflammatory lipid mediators by the inflammasome *in vivo*. *Nature* **490**, 107–111 (2012).
- Yang, J., Zhao, Y., Shi, J. & Shao, F. Human NALP and mouse NALP1 recognize bacterial type III secretion needle protein for inflammasome activation. *Proc. Natl Acad. Sci. USA* **110**, 14408–14413 (2013).
- Akhter, A. *et al.* Caspase-7 activation by the Nlrp4/Ipaf inflammasome restricts *Legionella pneumophila* infection. *PLoS Pathog.* **5**, e1000361 (2009).
- Sun, L. & Wang, X. A new kind of cell suicide: mechanisms and functions of programmed necrosis. *Trends Biochem. Sci.* **39**, 587–593 (2014).
- Poreba, M., Strozzyk, A., Salvesen, G. S. & Drag, M. Caspase substrates and inhibitors. *Cold Spring Harb. Perspect. Biol.* **5**, a008680 (2013).
- Fujii, T. *et al.* Gasdermin D (*Gsdmd*) is dispensable for mouse intestinal epithelium development. *Genesis* **46**, 418–423 (2008).
- Tanaka, S., Mizushima, Y., Kato, Y., Tamura, M. & Shiroishi, T. Functional conservation of *Gsdma* cluster genes specifically duplicated in the mouse genome. *G3 (Bethesda)* **3**, 1843–1850 (2013).
- Saeki, N. & Sasaki, H. in *Endothelium and epithelium: composition, functions, and pathology* (eds J. Carrasco & M. Matheus) Ch. IX 193–211 (Nova Science Publishers, 2011).
- Kumar, S. *et al.* *Gsdma3*^{359N} is a novel ENU-induced mutant mouse line for studying the function of Gasdermin A3 in the hair follicle and epidermis. *J. Dermatol. Sci.* **67**, 190–192 (2012).
- Ruge, F. *et al.* Delineating immune-mediated mechanisms underlying hair follicle destruction in the mouse mutant defolliculated. *J. Invest. Dermatol.* **131**, 572–579 (2011).
- Zhou, Y. *et al.* *Gsdma3* mutation causes bulge stem cell depletion and alopecia mediated by skin inflammation. *Am. J. Pathol.* **180**, 763–774 (2012).
- Shi, P. *et al.* Loss of conserved *Gsdma3* self-regulation causes autophagy and cell death. *Biochem. J.* **468**, 325–336 (2015).
- Agard, N. J., Maltby, D. & Wells, J. A. Inflammatory stimuli regulate caspase substrate profiles. *Mol. Cell. Proteomics* **9**, 880–893 (2010).
- Crawford, E. D. *et al.* The DegraBase: a database of proteolysis in healthy and apoptotic human cells. *Mol. Cell. Proteomics* **12**, 813–824 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank J. Ding for recombinant protein purification and X. Wang for reagents. We thank members of the Shao laboratory for helpful discussions and technical assistance. This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB08020202), the China National Science Foundation Program for Distinguished Young Scholars (31225002) and Program for International Collaborations (31461143006), and the National Basic Research Program of China 973 Program (2012CB518700 and 2014CB849602) to F.S. The research was supported in part by an International Early Career Scientist grant from the Howard Hughes Medical Institute and the Beijing Scholar Program to F.S.

Author Contributions F.S. and J.S. conceived the study; J.S. performed the CRISPR-Cas9 screens; J.S. and Y.Zha. designed and performed the majority of experiments, assisted by K.W. and X. S.; H.H. and T.C. performed the deep sequencing; J.S., Y.W., Y.Zhu. and F.W. generated the knockout mice. J.S., Y.Zha. and F.S. analysed the data and wrote the manuscript. All authors discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to F.S. (shaofeng@nibs.ac.cn).

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Plasmids, antibodies and reagents. Complementary DNA (cDNA) for human *GSDMD* was amplified from reverse-transcribed cDNA of HT-29 cells; cDNAs for human *GSDMB*, human *GSDMC* and mouse *Gsdma3* were synthesized by our in-house gene synthesis facility; cDNAs for human *GSDMA* and mouse *Gsdmd* were obtained from Vigene Biosciences (CH892815) and OriGene (MC202215), respectively. The gasdermin cDNAs were inserted into a modified pCS2-3×Flag vector for transient expression in 293T cells and the pWPI lentiviral vector with an N-terminal 2×Flag–HA tag or the FUIGW vector with an N-terminal Flag tag for stable expression in HeLa and iBMDM cells. For recombinant expression in *E. coli*, the cDNAs were cloned into a modified pET vector with an N-terminal SUMO tag. Truncation mutants of the gasdermins were constructed by the standard PCR cloning strategy and inserted into the pCS2 vector with indicated tags. Expression plasmids for caspase-1, 4, 5 and 11 were previously described^{4,9}, the caspase-9 plasmid was a gift from X. Wang (National Institute of Biological Sciences, Beijing). cDNAs for human *CASP2* and mouse *Casp8* are from the Life Technologies Ultimate ORF collection and OriGene (MC200404), respectively. Point mutations were generated by the QuickChange Site-Directed Mutagenesis Kit (Stratagene). All plasmids were verified by DNA sequencing.

Antibodies for caspase-1 p10 (sc-515), Myc epitope (sc-789) and *GSDMD* (sc-81868) were obtained from Santa Cruz Biotechnology. Other antibodies used in this study include anti-HA (MMS-101P, Covance), anti-Flag M2 (F4049), anti-actin (A2066) and anti-tubulin (T5168) (Sigma-Aldrich), rat monoclonal caspase-11 17D9 (NB120-10454, Novus Biologicals), anti-caspase-3 (#9662) and caspase-7 (#12827) (Cell Signaling Technology), IL-1β (3ZD; Biological Resources Branch, National Cancer Institute) and the antibody for detecting endogenous *GSDMD* (NBP2-33422, Novus Biologicals). Ultrapure LPS from *E. coli* O111:B4 and poly(dA:dT) were purchased from InvivoGen. LPS (L4524, for priming), TNFα and cycloheximide were purchased from Sigma-Aldrich. SMAC mimetic and the pan-caspase inhibitor zVAD are gifts from the laboratory of X. Wang (National Institute of Biological Sciences, Beijing). Nigericin was purchased from Calbiochem. Recombinant p20/p10 active caspase proteins (caspase-1/2/4/8/9) and lipid A (ALX-581-200-L001) were obtained from Enzo Life Sciences. Cell culture products are from Life technologies and all other chemicals used are Sigma-Aldrich products unless noted.

Cell culture and transfection. HeLa, HT-29 and 293T cells were obtained from ATCC. C57BL/6 mice-derived wild-type and *Tlr4*^{−/−} iBMDM cells were kindly provided by K. A. Fitzgerald (University of Massachusetts Medical School, United States) and A. Ding (Weill Cornell Medical College, United States), respectively, and used in our previous studies^{4,6,30}. All the cell lines are well-established, commonly used and frequently checked by virtue of their morphological features and functionalities, but have not been subjected to authentication by short tandem repeat (STR) profiling. All the cell lines have been tested to be mycoplasma-negative by the commonly used PCR method. iBMDM, HeLa and 293T cells were grown in Dulbecco's modified Eagle's medium (DMEM); HT-29 cells were grown in McCoy's 5a modified medium. All media were supplemented with 10% (vol/vol) fetal bovine serum (FBS) and 2 mM L-glutamine. All cells were grown at 37 °C in a 5% CO₂ incubator. Transient transfection of HeLa and 293T cells was performed using the JetPRIME (Polyplus Transfection) or Vigofect (Vigorous) reagents by following the manufacturers' instructions. For stable expression, lentiviral plasmids harbouring the desired gene were first transfected into 293T cells together with the packing plasmids pSPAX2 and pMD2G with a ratio of 5:3:2. The supernatants were collected 48 h after transfection and used to infect HeLa or iBMDM cells for another 48 h. GFP-positive infected cells were sorted by flow cytometry (BD Biosciences FACSARIA II). For siRNA knockdown, 0.5 μl of 20 μM siRNA together with 0.8 μl of INTERFERin reagents (Polyplus Transfection) were used for reverse transfection of iBMDM cells in the 96-well plate format; 5 μl of 20 μM siRNA and 10 μl of INTERFERin reagents were used to transfect HeLa cells in the 6-well plate format. The knockdown was performed for 60 h before subsequent analyses. The knockdown efficiency was assessed by quantitative real-time PCR (qRT-PCR) analyses as previously described⁴. All siRNA oligonucleotides were synthesized by our in-house facility using the sequences from the MISSION shRNA library (Broad Institute, United States) and their sequences are listed in Supplementary Table 1.

Inflammasome activation assays. Activation of the canonical caspase-1 inflammasomes (the NLRP3, NAIP–NLRC4 and AIM2 inflammasomes) and the non-canonical caspase-11 inflammasome by LPS was performed using the protocols that have been detailed in our previous publications^{4,6,9,20}. For bacteria-induced inflammasome activation, *S. typhimurium* (wild-type and *ΔsipD*), *B. thailandensis*

(wild-type and *ΔbipB*), EPEC (wild-type and *ΔescN*) were used to infect iBMDM cells and *S. typhimurium* (wild-type and *ΔsipA*) was used to infect HeLa cells, as described previously^{4,9}.

Microscopy imaging of cell death. To examine cell death morphology, cells were treated as indicated in the 6-well plates (Nunc products, Thermo Fisher Scientific Inc.) for static image capture or in glass-bottom culture dishes (MatTek Corporation) for live imaging. Static bright field images of pyroptotic cells were captured using an Olympus IX71 or a Zeiss Pascal Confocal microscope. The image pictures were processed using ImageJ or the LSM Image Examiner program. Live images of cell death were recorded with the PerkinElmer UltraVIEW spinning disk confocal microscopy and processed in the software Volocity. All image data shown were representative of at least three randomly selected fields.

Genome-wide CRISPR-Cas9 screens. The lentiviral gRNA plasmid library for genome-wide CRISPR-Cas9 screen was obtained from Addgene (#50947)³³; amplification of the library and preparation of the lentivirus were performed following the protocol provided by Addgene. In brief, 1 μl of library DNA (10 ng μl^{−1}) was used to transform 25 μl of electrocompetent *E. coli* (TaKaRa). Transformed colonies (>6 × 10⁷) were scraped off the Luria-Bertani (LB) plates into the media, and plasmids were extracted by using the GoldHi EndoFree Plasmid Maxi Kit (CWBIO). To prepare the virus library, 293T cells in the 15-cm dish were transfected with 25 μg of library DNA together with 15 μg of psPAX2 and 10 μg of pMD2G. Eight hours after transfection, the media were changed to high-serum DMEM (20% FBS with 25 mM HEPES). Another 40 h later, the media (from twenty 15-cm dishes of transfected cells) were collected and centrifuged at 3,000 r.p.m. for 10 min. The supernatant was filtered through a 0.22-μm membrane and aliquots of 30 ml were stored at −80 °C.

In the pilot experiment, the volume of the lentivirus library required for achieving an MOI of 0.3 for infecting the target cell line was determined in the 12-well plate format. For the large scale screen, *Tlr4*^{−/−} iBMDM cells stably expressing the Cas9 protein were seeded in the 15-cm dish (2 × 10⁶ cells in 20 ml media per dish) and a total of 2 × 10⁷ cells were infected with the gRNA lentivirus library. Sixty hours after infection, cells were re-seeded at a density of 1 × 10⁵ ml^{−1} in fresh media supplemented with 5 μg ml^{−1} puromycin (to eliminate non-infected cells). After 6 to 8 days, ~3 × 10⁸ cells from five culture dishes were electroporated with LPS to trigger caspase-11-mediated pyroptosis⁹, or stimulated with LFn–BsaK/protective antigen (PA) to induce caspase-1-mediated pyroptosis⁴; another 3 × 10⁸ cells were left untreated as the control sample. Each screen was repeated another time. Surviving cells were collected after growing to near 90% confluence and lysed in the SNET buffer (20 mM Tris-HCL (pH 8.0), 5 mM EDTA, 400 mM NaCl, 400 μg ml^{−1} Proteinase K and 1% SDS). Genomic DNAs of each group of cells were prepared by using the phenol-chloroform extraction and isopropanol precipitation method. The DNA was dissolved in H₂O (4–5 μg μl^{−1}) and used as the templates for amplification of the gRNA.

The gRNAs were amplified by a two-step PCR method using the Titanium Taq DNA polymerase (Clontech Laboratories). In the first step, six 50-μl PCR reactions (each containing 50 μg of genomic DNA template) were performed with the forward primer 50bp-F and the reverse primer 50bp-R; the PCR program used is 94 °C for 180 s, 16 cycles of 94 °C for 30 s, 60 °C for 10 s and 72 °C for 25 s, and a final 2-min extension at 68 °C. Products of the first-step PCR were pooled together and used as the template for the second-step PCR. Also six 50-μl PCR reactions (each containing 1 μl of the first-step PCR product) were performed with the forward primer Index-F and one of the reverse primers (Index-R1 to R6): Index-R1 for the control sample, Index-R2 for the replicate control sample, Index-R3 for the caspase-11 screen, Index-R4 for the replicate caspase-11 screen, Index-R5 for the caspase-1 screen and Index-R6 for the replicate caspase-1 screen. The PCR program used is 94 °C for 180 s, 18 cycles of 94 °C for 30 s, 54 °C for 10 s and 72 °C for 18 s, and a final 2-min extension at 68 °C. Products of the second-step PCR reactions were subjected to electrophoresis on the 1.5% agarose gel; the DNAs (the 310-bp band) were extracted and sequenced at the HiSeq2500 instrument (Illumina) by using the 50-bp single-end sequencing protocol. The first 19 nucleotides from each sequencing read are the gRNA sequence recovered from the library. The frequency of each gRNA was obtained by dividing the gRNA read number by the total sample read number; the fold of enrichment was calculated by comparing the frequency of each gRNA in the experiment sample with that in the control sample. Sequences for all the primers are listed in Supplementary Table 1.

The top 50 gRNA hits from the caspase-11 screen were examined and 18 genes that are conserved in human and mouse were identified for siRNA knockdown validation in HeLa cells. HeLa cells expressed caspase-4 but not caspase-5 (Extended Data Fig. 1b) and respond robustly to cytosolic LPS^{9,10}. For each gene, a mixture of two independent siRNAs was used and the knockdown efficiency of 12 of those having mRNA expression in HeLa cells was confirmed. Importantly, only siRNAs targeting human *GSDMD*, besides the control *CASP4*-targeting

siRNA, could efficiently block cytosolic LPS-induced pyroptosis (Extended Data Fig. 1c). When assayed individually, the two *GSDMD*-targeting siRNAs both showed potent inhibition of HeLa cell pyroptosis (Extended Data Fig. 1d).

Generation of CRISPR-Cas9 knockout cell lines. Human codon-optimized Cas9 (hCas9) and GFP-targeting gRNA-expressing plasmids (gRNA_GFP-T1) were purchased from Addgene. The 19-bp GFP-targeting sequence in the gRNA vector was replaced with the sequence targeting the desired gene by QuickChange site-directed mutagenesis. The target sequences used are AGCATCCTGGC ATTCCGAG for mouse *Gsdmd* and TTCCACTTCTACGATGCCA for human *GSDMD*. To construct the knockout cell lines, 1 µg of gRNA-expressing plasmid, 3 µg of hCas9 plasmid and 1 µg of pEGFP-C1 vector were co-transfected into 6×10^6 iBMDM or HeLa cells. Three days later, GFP-positive cells were sorted into single clones into the 96-well plate by flow cytometry using the BD Biosciences FACSaria II or the Beckman Coulter MoFlo XDP cell sorter. Single clones were screened by the T7 endonuclease I-cutting assay and the candidate knockout clones were verified by sequencing of the PCR fragments as described previously⁹. The PCR primers used are listed in Supplementary Table 1.

Knockout mice and primary BMDM cells. All animal experiments were conducted following the Ministry of Health national guidelines for housing and care of laboratory animals and performed in accordance with institutional regulations after review and approval by the Institutional Animal Care and Use Committee at National Institute of Biological Sciences. The *Gsdmd* knockout mice were generated by co-microinjection of *in vitro*-translated Cas9 mRNA and gRNA into the C57BL/6 zygotes. Founders with frameshift mutations were screened with T7E1 assay and validated by DNA sequencing. Founders were intercrossed to generate biallelic *Gsdmd*^{-/-} mice. The gRNA sequence used to generate the knockout mice is AGCATCCTGGCATTCCGAG. C57BL/6 wild-type mice were from Vital River Laboratory Animal Technology Co. and *Casp1/11*^{-/-} mice were obtained from the Jackson Laboratory. *Ripk3*^{-/-} mice were a gift from X. Wang (National Institute of Biological Sciences, Beijing). Primary BMDM cells were prepared from 6-week-old male mice (C57BL/6 background) by following a standard procedure as previously described⁶. For each experimental design, at least two mice were chosen to prepare the BMDM cells for assaying the inflammasome responses; the mice were not randomized and the investigators were not blinded.

Cytotoxicity assay and IL-1β ELISA. Relevant cells were treated as indicated. Cell death was measured by the LDH assay using CytoTox 96 Non-Radioactive Cytotoxicity Assay kit (Promega). Cell viability was determined by the CellTiter-Glo Luminescent Cell Viability Assay (Promega). To measure IL-1β release, primary BMDM cells were primed with LPS (1 µg ml⁻¹) for 2 h and released mature IL-1β was determined by using the IL-1β ELISA kit (Neobioscience Technology Company).

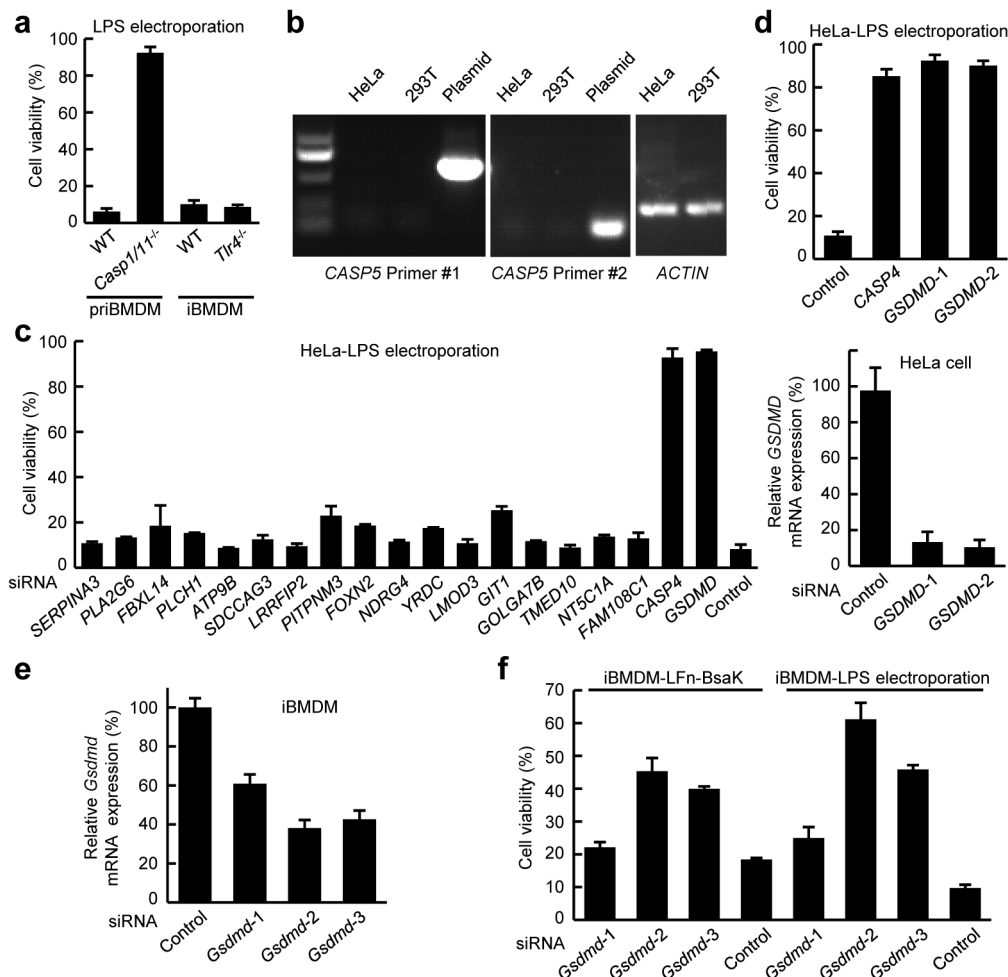
Purification of recombinant proteins. To obtain recombinant human GSDMD, *E. coli* BL21 (DE3) cells harbouring pET28a-His₆-SUMO-GSDMD were grown in LB medium supplemented with 30 µg ml⁻¹ kanamycin. Protein expression was induced overnight at 18 °C with 0.4 mM isopropyl-B-D-thiogalactopyranoside (IPTG) after OD_{600 nm} reached 0.8. Cells were harvested and resuspended in a

lysis buffer containing 20 mM Tris-HCl (pH 8.0), 150 mM NaCl, 20 mM imidazole and 10 mM 2-mercaptoethanol. The His₆-SUMO-tagged protein was first purified by affinity chromatography using Ni-NTA beads (Qiagen) and the SUMO tag was removed by overnight ULP1 protease digestion at 4 °C. The cleaved GSDMD was further purified by HiTrap Q ion-exchange and Superdex G200 gel-filtration chromatography (GE Healthcare Life Sciences).

To obtain the constitutive-active caspase-11 p20/p10 tetramer, cDNAs encoding the p20 large and p10 small subunit were cloned into pET21a with a 6×His tag fused to the C terminus of the p10 subunit. The two subunits were separately expressed in *E. coli* with 1 mM IPTG induction for 4 h at 30 °C. Bacteria collected from 1-l culture were resuspended and lysed in 100 ml of lysis buffer (50 mM Tris-HCl (pH 8.0), 150 mM NaCl and 10 mM 2-mercaptoethanol) by sonication. Inclusion bodies, obtained by centrifugation of the lysates at 18,000 r.p.m. for 1 h, was washed with 50 ml of Buffer 1 (50 mM Tris-HCl (pH 8.0), 300 mM NaCl, 1 M guanidinium hydrochloride (GdnCl) and 0.1% Triton X-100) and 50 ml of Buffer 2 (50 mM Tris-HCl (pH 8.0), 300 mM NaCl and 1 M GdnCl) twice for each buffer. The washed inclusion bodies were solubilized by stirring in 6 ml of the solubilization buffer containing 6.5 M GdnCl, 25 mM Tris-HCl (pH 7.5), 5 mM EDTA and 100 mM DTT overnight at room temperature. To obtain active p20/p10 tetramers by refolding, 12 ml of above solubilized inclusion body solution containing denatured large and small subunits (molecular ratio, 1:2) were drop-by-drop diluted in 500 ml of refolding buffer (100 mM HEPES, 100 mM NaCl, 100 mM sodium malonate, 20% sucrose, 0.1 M NDSB-201 and 10 mM DTT) and then gently stirred in a nitrogen atmosphere at 16 °C overnight. Protein aggregates were removed by centrifugation at 4,000 r.p.m. for 20 min and the refolded protein supernatants were concentrated and dialysed against a buffer containing 50 mM Tris-HCl (pH 8.0), 150 mM NaCl and 10 mM 2-mercaptoethanol. The protein was affinity-purified by the Ni-NTA beads and further purified by the Superdex G200 gel-filtration chromatography. Expression and purification of recombinant LFn-BsaK and LFn-FliC proteins were described previously⁴. Recombinant full-length caspase-4 and caspase-11 were expressed and purified from insect cells also as previously described⁹. Recombinant PreScission protease (PPase) proteins are routine lab stocks.

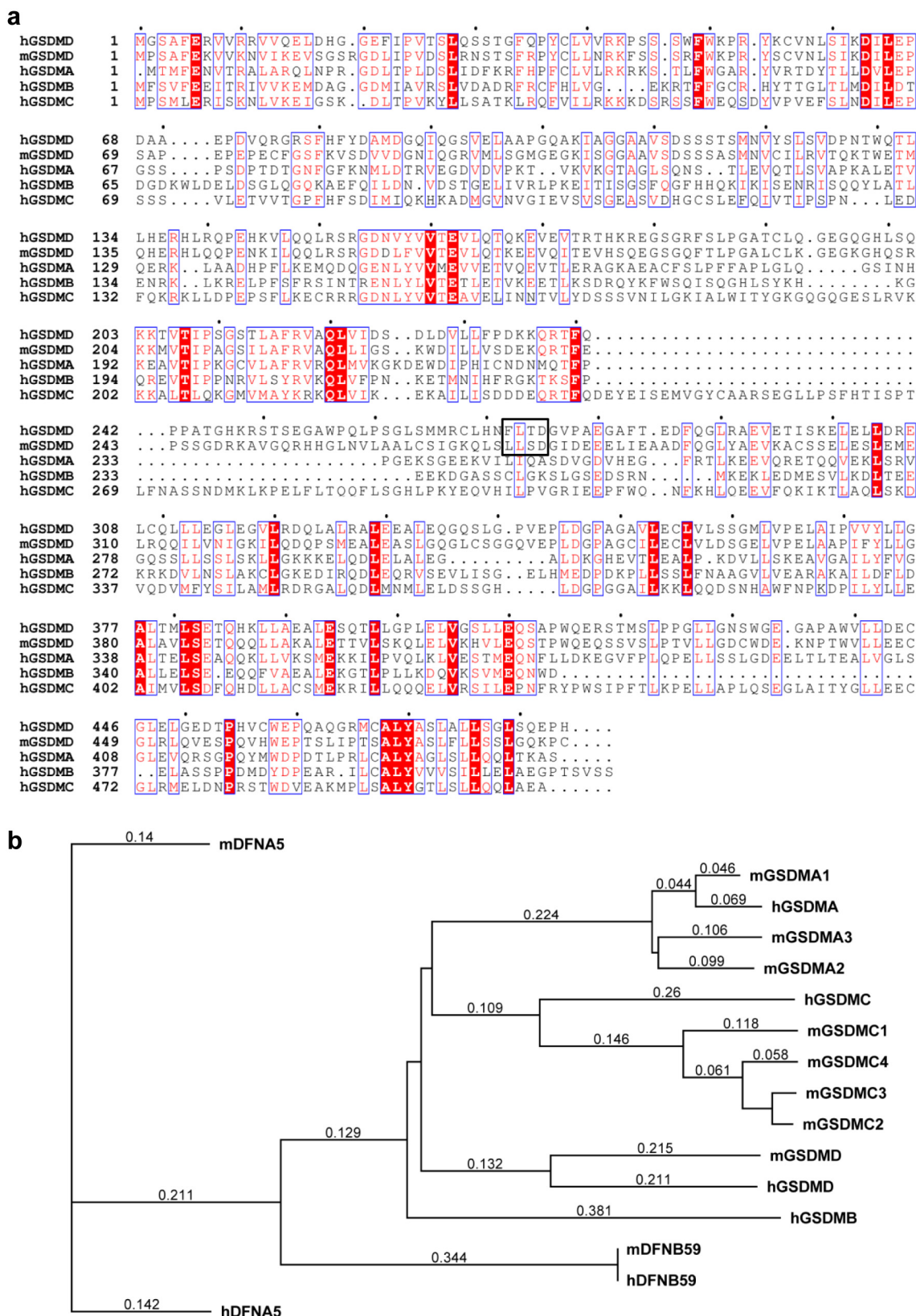
In vitro GSDMD cleavage by recombinant caspases. For cleavage by the p20/p10 tetramers of active caspase, 5 µg of purified recombinant GSDMD was incubated with 1 unit of caspase-1, 2, 4, 8 and 9 or 0.1 µg of caspase-11 in a 25-µl reaction containing 50 mM HEPES (pH 7.5), 3 mM EDTA, 150 mM NaCl, 0.005% (vol/vol) Tween-20 and 10 mM DTT. The reaction was incubated for 60 min at 37 °C. For cleavage by LPS-activated caspase-4/11, the full-length caspase proteins purified from insect cells were first incubated with LPS, lipid A or MDP for 30 min at 30 °C; 5 µg of purified recombinant GSDMD was then reacted with the ligand-incubated caspases at 37 °C for 9 min. Cleavage of GSDMD was examined by Coomassie blue staining of the reaction samples separated on the SDS-PAGE gel.

33. Koike-Yusa, H., Li, Y., Tan, E.-P., del Castillo Velasco-Herrera, M. & Yusa, K. Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. *Nature Biotechnol.* **32**, 267–273 (2014).



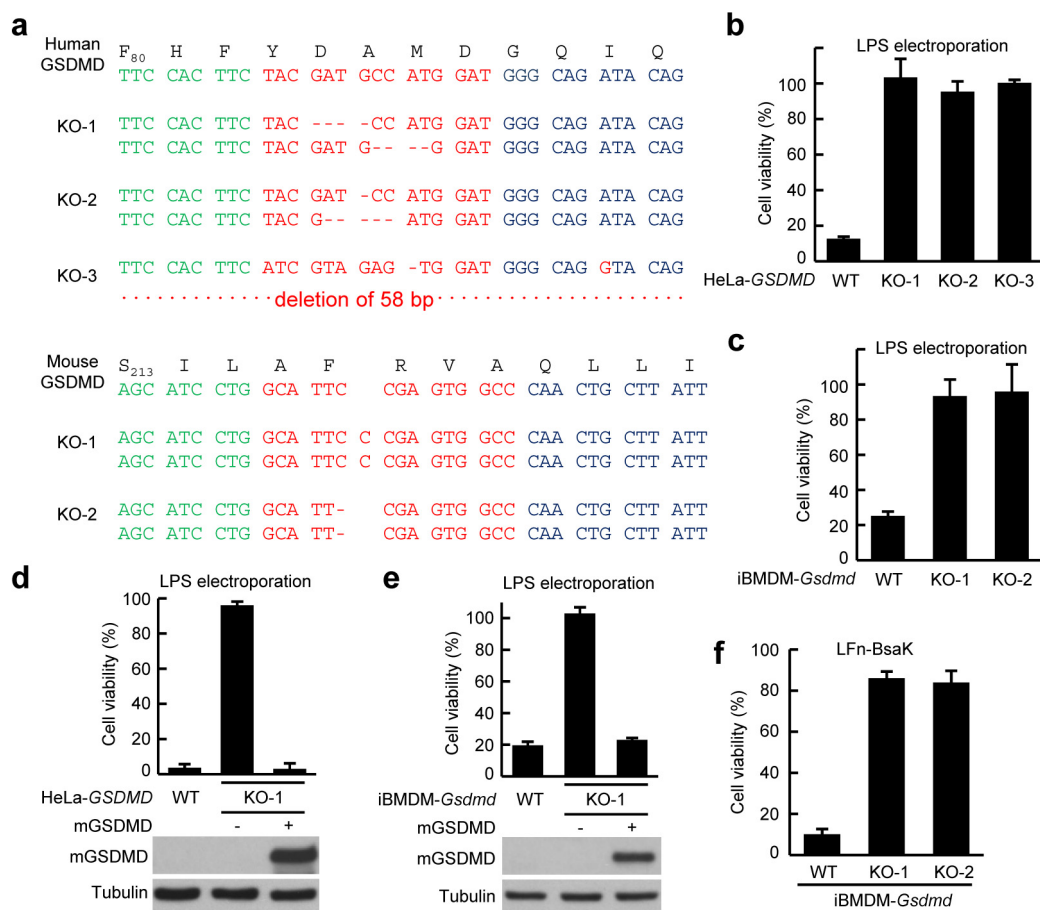
Extended Data Figure 1 | siRNA knockdown validation of the requirement of GSDMD for LPS- and LFn-BsaK-induced pyroptosis. **a**, LPS electroporation-induced pyroptosis in the absence of priming. Primary BMDMs (priBMDM) (wild-type (WT) or the *Casp1* and *Casp11* double knockout) or iBMDM cells (wild-type or *Tlr4*^{-/-}) were assayed. *Tlr4*^{-/-} iBMDMs were used for the CRISPR-Cas9 screen in this study. **b**, Reverse-transcription PCR analyses of caspase-5 expression in HeLa and 293T cells. Plasmid harbouring caspase-5 cDNA serves as the positive control. **c**, siRNA knockdown validation of the CRISPR-Cas9 screen of LPS-induced pyroptosis. HeLa cells were used to validate the selected top hits from the

screen. Mixtures of two independent siRNA pairs targeting each gene were transfected into the cells. siRNAs targeting *CASP4* and luciferase were used as the positive and negative control, respectively. **d**, Effects of *GSDMD* siRNA knockdown on LPS-induced pyroptosis in HeLa cells. **e**, **f**, Effects of *Gsdmd* siRNA knockdown on LPS and LFn-BsaK-induced pyroptosis in iBMDM cells. The knockdown efficiency (**d**, **e**) was measured by qRT-PCR analyses. ATP-based cell viability (**a**, **c**, **d**, **f**) and siRNA knockdown efficiency (**d**, **e**) were expressed as mean values \pm s.d. from three technical replicates. Data shown are representative of two (**c**) or three (**a**, **b**, **d**-**f**) independent experiments.



Extended Data Figure 2 | The gasdermin family of proteins in human and mouse. **a**, Multiple sequence alignment of human GSDMA, GSDMB, GSDMC, GSDMD and mouse GSDMD. The alignment was performed by using the ClustalW2 algorithm and displayed with ESPrnt 3.0 (<http://esprnt.ibcp.fr/ESPrnt/cgi-bin/ESPrnt.cgi>). Identical residues are highlighted by the dark red background and conserved residues are indicated by red font. The black box

marks the caspase-1/4/11 cleavage motifs in human and mouse GSDMD. The residue number is indicated on the left of the sequence. **b**, Phylogenetic tree of all the gasdermin family of proteins in human and mouse. ClustalW alignment was carried out to generate the phylogenetic tree by using the ‘Neighbor Joining’ method. DFNA5 and DFNB59 are distantly related to the gasdermins, and the latter only contains gasdermin-N domain.

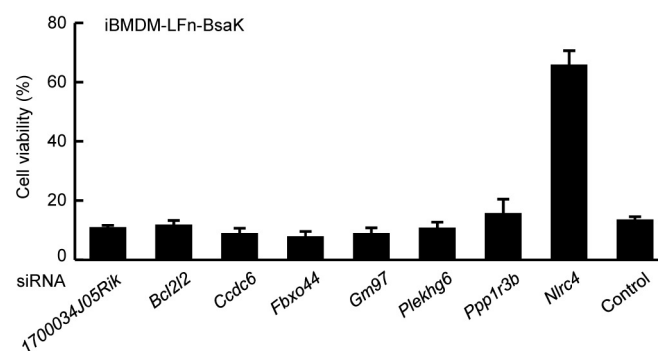
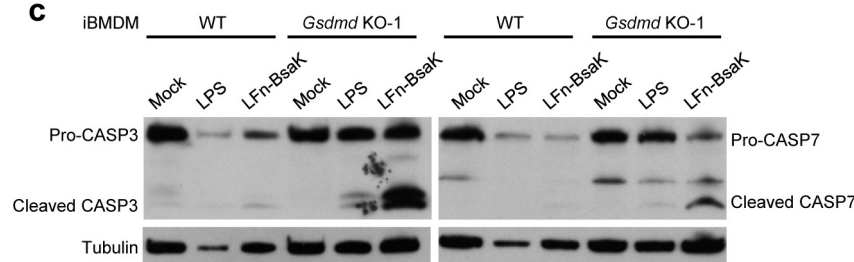


Extended Data Figure 3 | Generation of GSDMD-deficient cell lines and assays for cytosolic LPS- and LFn-BsaK-triggered pyroptosis. a, Generation of *GSDMD*^{-/-} HeLa cells and *Gsdmd*^{-/-} iBMDM cells by CRISPR-Cas9-mediated targeting. Shown are the sequence mutations of the three HeLa cell clones and two iBMDM clones used in the study. **b, c, f**, Effects of *GSDMD*^{-/-} on LPS electroporation-induced pyroptosis in HeLa (**b**) and iBMDM cells (**c**) and LFn-BsaK-induced pyroptosis in iBMDM cells (**f**).

d, e, Complementation of *GSDMD*^{-/-} HeLa cells and *Gsdmd*^{-/-} iBMDM cells by stably expressed mouse GSDMD-3×Flag. The accompanying blots show the expression of exogenous GSDMD by anti-Flag immunoblotting with anti-tubulin blots serving as the loading control. ATP-based cell viability is expressed as mean values ± s.d. from three technical replicates (**b–f**). Data shown are representative of at least three independent experiments.

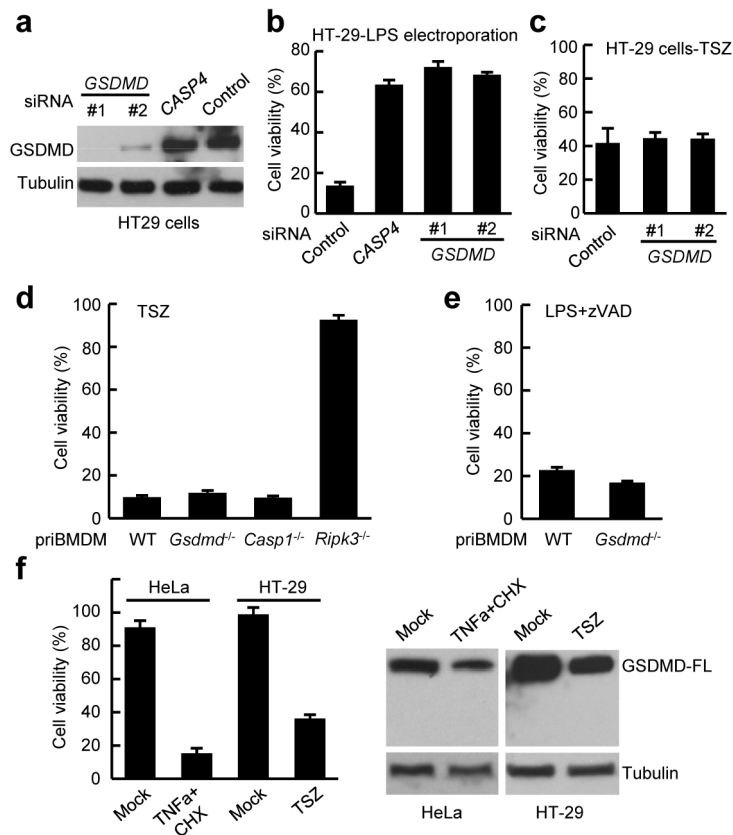
a LFn-BsaK CRISPR-Cas9 screen hits with multiple gRNAs

Ranking	Gene_gRNA #	Average folds of enrichment	gRNA sequence
1	<i>Nlrc4_2</i>	3695.249	GTTTCGAATAGTCCCCCCC
2	<i>Nlrc4_1</i>	2797.553	AGGTGCCTCATGACCGCCC
3	<i>Nlrc4_5</i>	2202.819	ATATGGATGATCTTTCGGG
4	<i>Nlrc4_4</i>	1798.824	ATCAGCAAGCCGACCTTCA
6	<i>Naip2_5</i>	1599.965	TTCTGGTCGTAATTTCTTG
7	<i>Nlrc4_3</i>	1241.855	ATCTGCTCCTCTACACGTG
8	<i>Casp1_3</i>	1143.169	AGGGCAAGACGTGTACGAG
10	<i>Casp1_1</i>	877.4373	GAATTCTGGAGCTTCAATC
12	<i>Fbxo44_1</i>	659.6165	ATCGCCTCCGTTACGTCC
13	<i>Casp1_5</i>	649.6994	ATGTCTCATGGTATCCAGG
15	<i>Naip2_4</i>	554.2724	ACTGGCCCCACGAATCACC
17	<i>Naip2_2</i>	315.3426	ACAGCCCCGGGTGATTCTG
18	<i>Casp1_4</i>	246.0405	CAACTTGAGCTCCAACCCT
21	<i>Naip2_1</i>	202.6801	CTAGACTCGTATCTAGGTA
36	<i>Antxr2_4</i>	67.18251	TGACGGACGGTAAGCTGGA
57	<i>Ccdc6_3</i>	35.80221	CTCGCCGTGAATTACGAGA
88	<i>Ccdc6_5</i>	23.54041	CTCGCCGTGAATTACGAGA
138	<i>Gsdmd_3</i>	16.5958	CAACAGCTTCGGAGTCGTG
153	<i>1700034J05Rik_4</i>	15.40841	CTCTTCAGGAAACGAGATC
171	<i>Plekhg6_3</i>	14.6932	CAGACCATGGCTTATGCGC
195	<i>Bcl2l2_1</i>	13.87979	CCTAACTGGGGCCGCTTTG
203	<i>Plekhg6_1</i>	13.46514	ATAAATGGCCAGGTCCGAC
236	<i>Fbxo44_4</i>	12.63271	GGCTTCATCACCAGGAGCT
239	<i>Antxr2_2</i>	12.47934	GCTAGTGTCTTACTGCGTTG
338	<i>Bcl2l2_5</i>	9.98516	CCTAACTGGGGCCGCTTTG
340	<i>1700034J05Rik_2</i>	9.970983	CTGGCCTGGGATATCGATG
352	<i>Gsdmd_4</i>	9.753432	AGCATCCTGGCATTCCGAG
353	<i>Gm97_5</i>	9.727808	CTGGCTCACCTATATGCC
365	<i>Gm97_4</i>	9.490875	CTGGCTCACCTATATGCC
376	<i>Ppp1r3b_3</i>	9.245637	AGCCCCTGGTTGTCGGCGA
380	<i>Ppp1r3b_5</i>	9.168484	ACAGTTTCTAGGCAGACG

b**c**

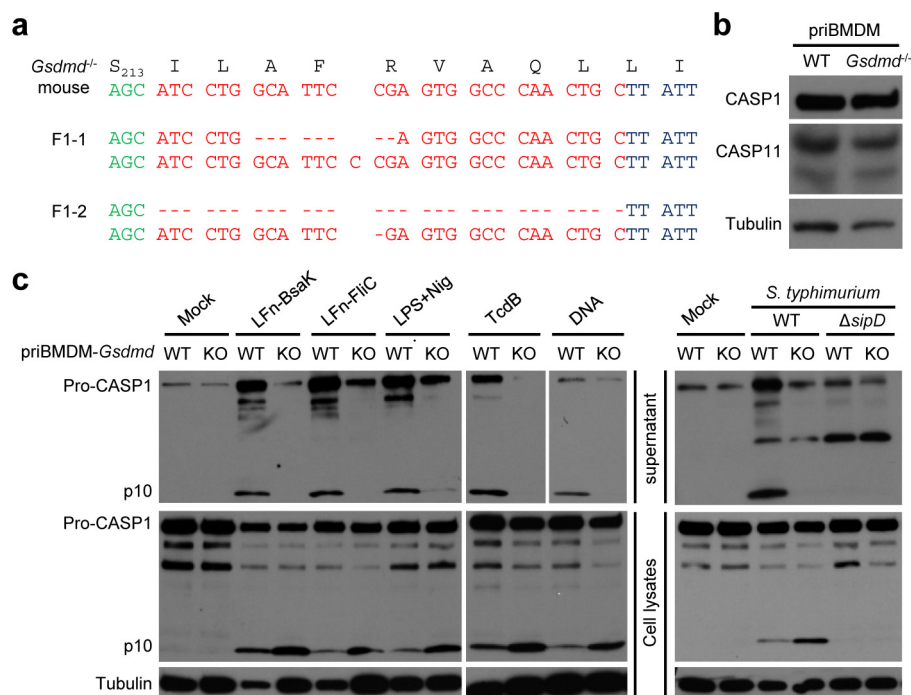
Extended Data Figure 4 | CRISPR-Cas9 screen of LFn-BsaK-triggered pyroptosis and effects of *Gsdmd* knockout on LFn-BsaK-induced and caspase-1-mediated caspase-3/7 cleavage. **a**, gRNA hits from a genome-wide CRISPR-Cas9 screen of LFn-BsaK-induced pyroptosis in mouse *Thr4^{-/-}* iBMDM cells. Shown are those genes with multiple gRNA hits. The ranking, the average fold increase and the sequences for each gRNA are listed. Genes highlighted in red encode known components in the pathway (*Antxr2* encodes the endocytosis receptor for the LFn tag) and were hit by multiple gRNAs. **b**, siRNA knockdown validation of screen hits. Mixtures of two independent

siRNA pairs targeting each gene were transfected into the iBMDM cells before stimulation with LFn-BsaK. siRNAs targeting *Nlrc4* and luciferase were used as the positive and negative control, respectively. ATP-based cell viability is expressed as mean values \pm s.d. from three technical replicates. **c**, Caspase-3/7 activation upon prolonged LFn-BsaK treatment in wild-type and *Gsdmd^{-/-}* (the KO-1 clone; KO, knockout) iBMDM cells. Cell lysates were analysed by anti-caspase-3/7 and tubulin immunoblotting. Data shown are representative of two (**b**) and three (**c**) independent experiments.



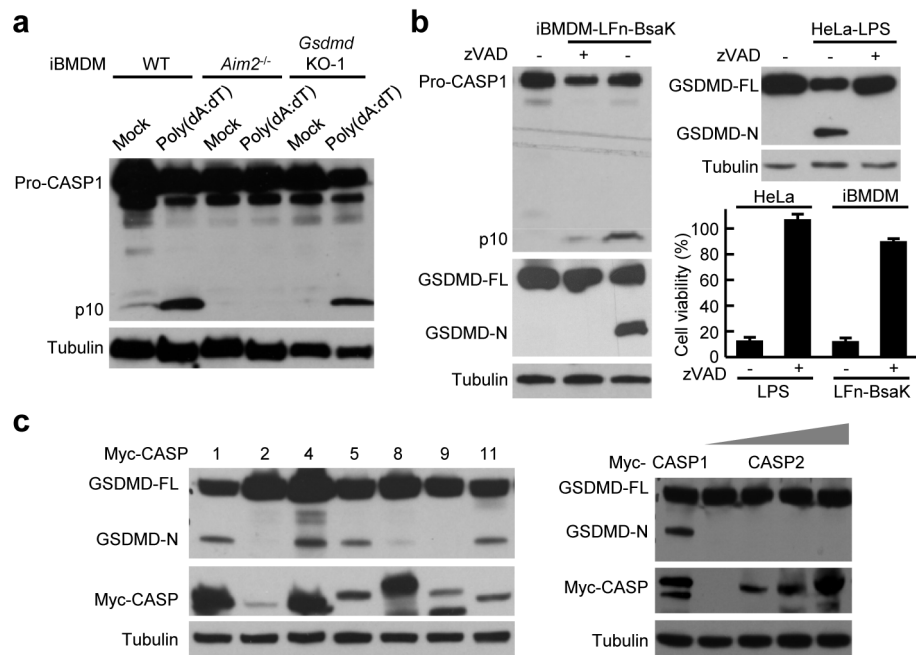
Extended Data Figure 5 | GSDMD is not required for and not cleaved in TNF α -induced necroptosis and apoptosis. **a–c**, Effects of *GSDMD* knockdown on LPS electroporation-induced pyroptosis and TSZ-induced necroptosis in HT-29 cells. Two independent *GSDMD*-targeting siRNAs (#1 and #2) were assayed and the immunoblots in **a** show the knockdown efficiency. **d, e**, Primary BMDM cells from *Gsdmd*^{-/-} or other indicated mouse strains were stimulated with TSZ (**d**) or LPS + zVAD (**e**) to trigger necroptosis. **f**, The absence of GSDMD cleavage in TNF α -induced apoptosis and

necroptosis. 2 \times Flag–HA–GSDMD was stably expressed in HeLa and HT-29 cells. Apoptosis was induced by TNF α + CHX treatment in HeLa cells and necroptosis was induced by TSZ stimulation of HT-29 cells. Lysates of stimulated cells were analysed by anti-Flag and anti-tubulin immunoblotting to examine possible GSDMD cleavage. GSDMD-FL, full-length GSDMD. ATP-based cell viability is expressed as mean values \pm s.d. from three technical replicates (**b–f**). Data shown are representative of at least two independent experiments.



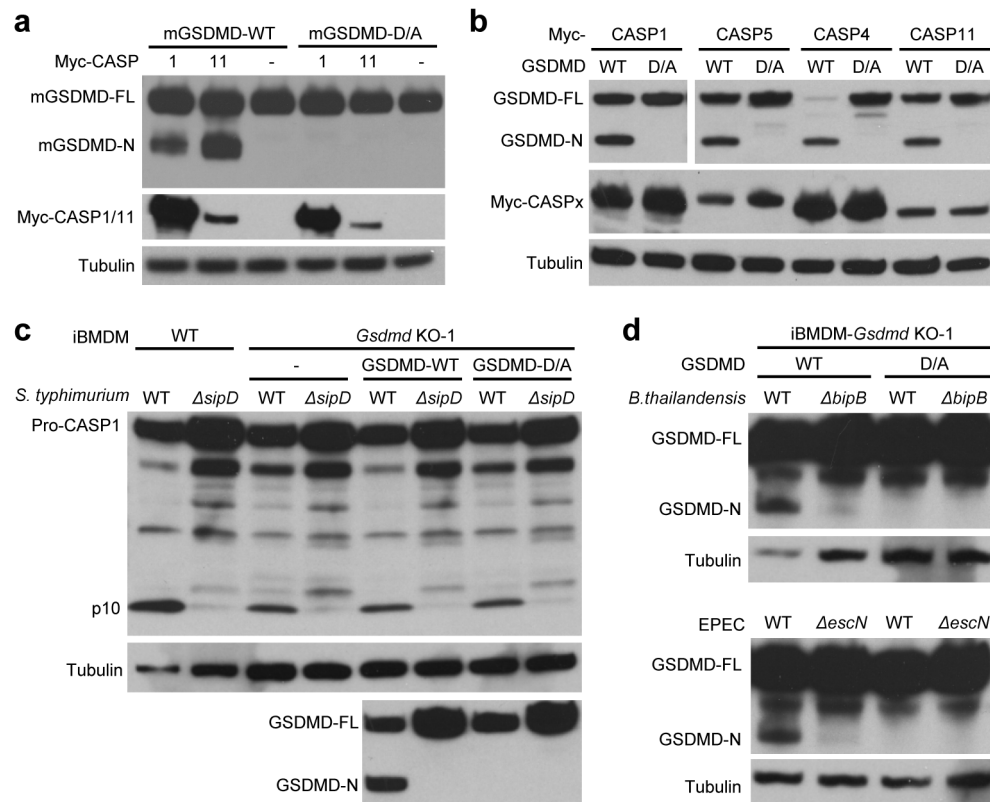
Extended Data Figure 6 | Generation of *Gsdmd*^{-/-} mice and assays for inflammasome-mediated caspase-1 autoprocessing and secretion.
a, *Gsdmd*^{-/-} mice were generated by CRISPR-Cas9-mediated targeting. Shown are the sequence mutations in the two homozygous F1 lines (F1-1 and F1-2) used in the study. **b**, Anti-caspase-1/caspase-11 immunoblots of lysates of unstimulated primary BMDM cells derived from wild-type and *Gsdmd*^{-/-}

mice. **c**, Primary BMDMs derived from wild-type or *Gsdmd*^{-/-} mice were stimulated with indicated canonical inflammasome stimuli or infected with *S. typhimurium* (wild type or the T3SS-deficient $\Delta sipD$ mutant). Total cell lysates or the culture supernatants were subjected to anti-caspase-1 or anti-tubulin immunoblotting. Data shown (**b**, **c**) are representative of two independent experiments.



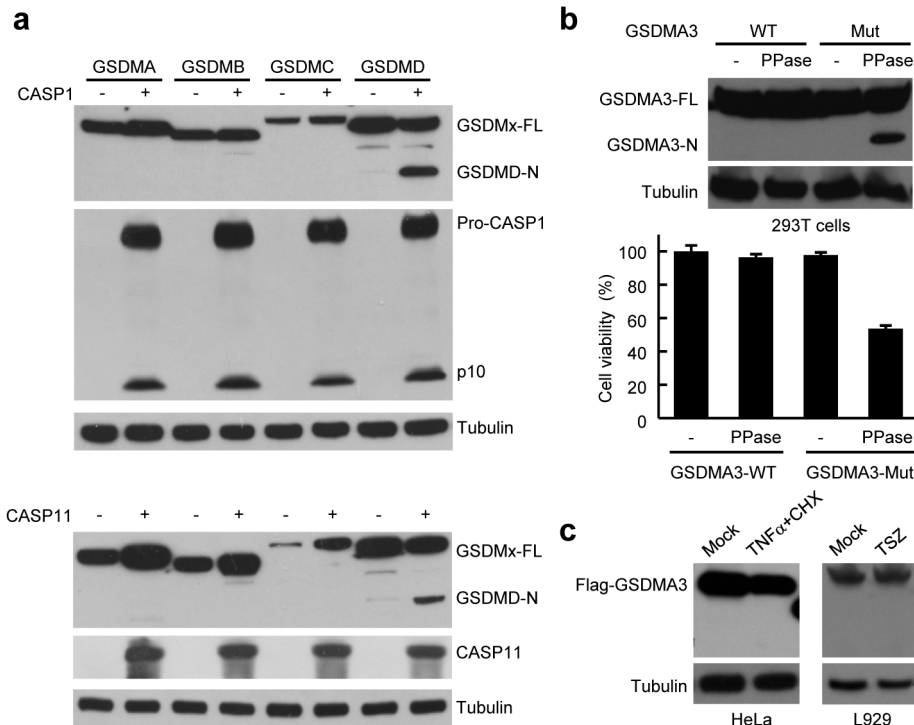
Extended Data Figure 7 | Specific cleavage of GSDMD by inflammatory caspases. **a**, Effects of *Gsdmd* knockout on caspase-1 activation by the AIM2 inflammasome. Indicated iBMDM cells were stimulated by poly(dA:dT) transfection. **b**, Effects of the pan-caspase inhibitor zVAD on LPS electroporation- and LFn-BsaK-induced GSDMD cleavage in HeLa and iBMDM cells, respectively. ATP-based cell viability is expressed as mean values

± s.d. from three technical replicates. **c**, Assays of GSDMD cleavage by inflammatory and apoptotic caspases overexpressed in cells. 3×Flag-GSDMD was co-transfected with indicated Myc-caspase into 293T cells. Total cell lysates were analysed by anti-caspase-1 (**a**, **b**), anti-Flag (**b**, **c**), anti-Myc (**c**) and anti-tubulin (**a**–**c**) immunoblotting. Data shown are representative of three independent experiments.



Extended Data Figure 8 | Resistance of the GSDMD D/A mutant to inflammatory-caspase cleavage. **a, b**, Assays of proteolytic cleavage of the GSDMD D/A mutant by overexpression-activated inflammatory caspases. 3×Flag-tagged mouse (**a**) or human (**b**) GSDMD (wild-type or the D/A mutant) was co-transfected with Myc-tagged caspase-1/11 (**a**) or caspase1/4/5/11 (**b**) into 293T cells. Cell lysates were analysed by anti-Flag, anti-Myc and anti-tubulin immunoblotting. **c, d**, Assays of proteolytic cleavage of GSDMD D/A mutant by bacterial-infection-activated caspase-1. Wild-type, *Gsdmd* knockout (the KO-1 clone), or *Gsdmd* KO-1 complemented with 2×Flag-

HA-GSDMD (wild-type or the D/A mutant) iBMDM cells were infected with wild-type *S. typhimurium* (**c**), *B. thailandensis* or EPEC (**d**) to induce caspase-1 activation (by the NAIP-NLRC4 inflammasome), or their T3SS-deficient mutant strains ($\Delta sipD$, $\Delta bipB$ and $\Delta escN$, respectively) as controls. Cell lysates were analysed by anti-caspase-1, anti-tubulin and anti-Flag immunoblotting. p10, mature caspase-1. GSDMD-FL, full-length GSDMD; GSDMD-N, the N-terminal cleavage product of GSDMD. The D/A mutants refer to D275A for human GSDMD and D276A for mouse GSDMD. Data shown are representative of three independent experiments.



Extended Data Figure 9 | Characterization of GSDMA3 and other gasdermin family members. **a**, Flag-tagged GSDMA, GSDMB, GSDMC and GSDMD were co-transfected with caspase-1 (upper panel) or caspase-11 (lower panel) into 293T cells. Cell lysates were analysed by anti-Flag, anti-caspase-1, anti-Myc or anti-tubulin immunoblotting. **b**, Wild-type GSDMA3 or a GSDMA3-mutant harbouring a PPase cleavage site between its gasdermin-N and -C domain was expressed in 293T cells. Recombinant PPase was transfected into the cells by electroporation. The upper panel shows the

immunoblots of cell lysates to examine GSDMA3 cleavage and the lower panel shows ATP-based cell viability expressed as mean values \pm s.d. from three technical replicates. **c**, The absence of GSDMA3 cleavage in TNF α -induced apoptosis and necroptosis. Flag-GSDMA3 was expressed in HeLa and L929 cells. Apoptosis was induced by TNF α + CHX treatment in HeLa cells and necroptosis was induced by TSZ stimulation of L929 cells. Lysates of stimulated cells were analysed by anti-Flag and anti-tubulin immunoblotting. Data shown are representative of three independent experiments.

Caspase-11 cleaves gasdermin D for non-canonical inflammasome signalling

Nobuhiko Kayagaki¹, Irma B. Stowe¹, Bettina L. Lee¹, Karen O'Rourke¹, Keith Anderson², Søren Warming², Trinna Cuellar², Benjamin Haley³, Merone Roose-Girma², Qui T. Phung³, Peter S. Liu³, Jennie R. Lill³, Hong Li³, Jiansheng Wu³, Sarah Kummerfeld⁴, Juan Zhang⁵, Wyne P. Lee⁵, Scott J. Snipas⁶, Guy S. Salvesen⁶, Lucy X. Morris⁷, Linda Fitzgerald⁷, Yafei Zhang⁷, Edward M. Bertram^{7,8}, Christopher C. Goodnow^{8,9,10} & Vishva M. Dixit¹

Intracellular lipopolysaccharide from Gram-negative bacteria including *Escherichia coli*, *Salmonella typhimurium*, *Shigella flexneri*, and *Burkholderia thailandensis* activates mouse caspase-11, causing pyroptotic cell death, interleukin-1 β processing, and lethal septic shock. How caspase-11 executes these downstream signalling events is largely unknown. Here we show that gasdermin D is essential for caspase-11-dependent pyroptosis and interleukin-1 β maturation. A forward genetic screen with ethyl-*N*-nitrosourea-mutagenized mice links *Gsdmd* to the intracellular lipopolysaccharide response. Macrophages from *Gsdmd*^{-/-} mice generated by gene targeting also exhibit defective pyroptosis and interleukin-1 β secretion induced by cytoplasmic lipopolysaccharide or Gram-negative bacteria. In addition, *Gsdmd*^{-/-} mice are protected from a lethal dose of lipopolysaccharide. Mechanistically, caspase-11 cleaves gasdermin D, and the resulting amino-terminal fragment promotes both pyroptosis and NLRP3-dependent activation of caspase-1 in a cell-intrinsic manner. Our data identify gasdermin D as a critical target of caspase-11 and a key mediator of the host response against Gram-negative bacteria.

Cytoplasmic caspase-11 (also known as caspase-4) defines the non-canonical inflammasome that is activated by various Gram-negative bacterial infections and causes infected cells to die by pyroptosis¹⁻⁴. Caspase-11 also triggers NLRP3- and ASC-dependent activation of caspase-1, resulting in the proteolytic maturation of the inactive cytokine pro-interleukin (IL)-1 β (refs 1, 5–7). Lipopolysaccharide (LPS) is the bacterial pathogen-associated molecular pattern that promotes caspase-11 activation^{7,8}. Cytoplasmic LPS can directly bind to and activate caspase-11 independently of Toll-like receptor 4 (TLR4)⁹, but precisely how caspase-11 executes pyroptosis and non-canonical NLRP3 inflammasome activation has remained unknown.

Phenomics identifies gasdermin D

An unbiased forward genetic screen with ethyl-*N*-nitrosourea (ENU)-mutagenized mice¹⁰ was performed to unveil essential mediators of non-canonical inflammasome signalling in response to cytoplasmic LPS. Pedigree IGL1351 had a Mendelian recessive mutation that compromised LPS-induced IL-1 β secretion from peritoneal macrophages. The severity of the defect was equivalent to that seen with the *Casp11* mutant 129 strain¹ (Fig. 1a and Extended Data Fig. 1). Importantly, IL-1 β secretion by pedigree IGL1351 in response to *Clostridium difficile* toxin B (TcdB), a stimulant of the Mediterranean fever/Pyrim canonical inflammasome^{11,11}, was comparable to other pedigrees. Exome sequencing identified 39 single-nucleotide variants (SNVs) in the ENU-mutated founder G1 male. Subsequent genotyping of all identified SNVs in a large cohort of siblings and offspring revealed that the trait was completely correlated with inheritance of a point mutation in the gene encoding gasdermin D (*Gsdmd*) (Extended Data Fig. 1 and Extended Data Table 1).

Gsdmd (ref. 12) is a 487 amino acid cytoplasmic protein that contains an ill-characterized gasdermin domain and lacks any obvious signal

peptide or transmembrane segments (Fig. 1b). The human gasdermin gene family consists of *GSDMA*, *GSDMB*, *GSDMC* and *GSDMD*. The physiological role of *Gsdmd* is unknown¹³, although other gasdermin family members have been implicated in deafness or apoptosis in epithelial cells^{12,14}. The ENU-derived mutation in *Gsdmd* caused an Ile to Asn amino acid substitution at position 105 with a PolyPhen score of 0.98 (probably damaging) and a SIFT (sorting intolerant from tolerant) score of 0.00 (deleterious) (Fig. 1b and Extended Data Table 1). Full-length *Gsdmd* protein (53 kDa) was immunoblotted in unstimulated wild-type bone-marrow-derived macrophages (BMDMs). Unlike pro-caspase-11, *Gsdmd* was not induced to any great extent by TLR agonists or interferons (Fig. 1c). Wild-type and *Gsdmd*^{I105N/I105N} BMDMs expressed comparable amounts of *Gsdmd* (Fig. 1d), suggesting that the I105N mutation impairs *Gsdmd* function rather than protein expression.

Wild-type, *Casp11*^{-/-}, or *Gsdmd*^{I105N/I105N} BMDMs that were primed with Pam3CSK4 and then stimulated with canonical inflammasome stimuli, including ATP to engage the NLRP3-dependent inflammasome⁶, poly(dA:dT) double-stranded DNA (dsDNA) to activate the AIM2-dependent inflammasome^{15–17}, or TcdB¹¹, secreted comparable amounts of IL-1 β and exhibited similar levels of pyroptotic death¹ (Fig. 1e). Consistent with these data, *Gsdmd*^{I105N/I105N} BMDMs displayed normal induction of NLRP3, pro-IL-1 β , and RANTES cytokine in response to TLR agonists (Fig. 1d, Extended Data Fig. 2a). *Gsdmd*^{I105N/I105N} BMDMs also upregulated pro-caspase-11 normally (Fig. 1d) but, in contrast to their wild-type counterparts, were defective in IL-1 β secretion, and did not undergo pyroptosis in response to cytoplasmic LPS (Fig. 1e). Proteolytic processing of caspase-1 and pro-IL-1 β preceding IL-1 β secretion was also impaired, suggesting that *Gsdmd* is critical for all caspase-11-dependent signalling events (Extended Data Fig. 2b).

¹Department of Physiological Chemistry, Genentech Inc., South San Francisco, California 94080, USA. ²Department of Molecular Biology, Genentech Inc., South San Francisco, California 94080, USA.

³Department of Protein Chemistry, Genentech Inc., South San Francisco, California 94080, USA. ⁴Department of Bioinformatics, Genentech Inc., South San Francisco, California 94080, USA. ⁵Department of Immunology, Genentech Inc., South San Francisco, California 94080, USA. ⁶Program in Cell Death Signaling Networks, Sanford-Burnham-Prebys Medical Discovery Institute, La Jolla, California 92037, USA. ⁷The Australian Phenomics Facility, The John Curtin School of Medical Research, The Australian National University, Canberra, Australian Capital Territory 2601, Australia. ⁸Department of Immunology and Infectious Diseases, The John Curtin School of Medical Research, The Australian National University, Canberra, Australian Capital Territory 2601, Australia. ⁹Garvan Institute of Medical Research, Sydney, New South Wales 2010, Australia. ¹⁰St. Vincent's Clinical School, UNSW Australia, Darlinghurst, New South Wales 2010, Australia.

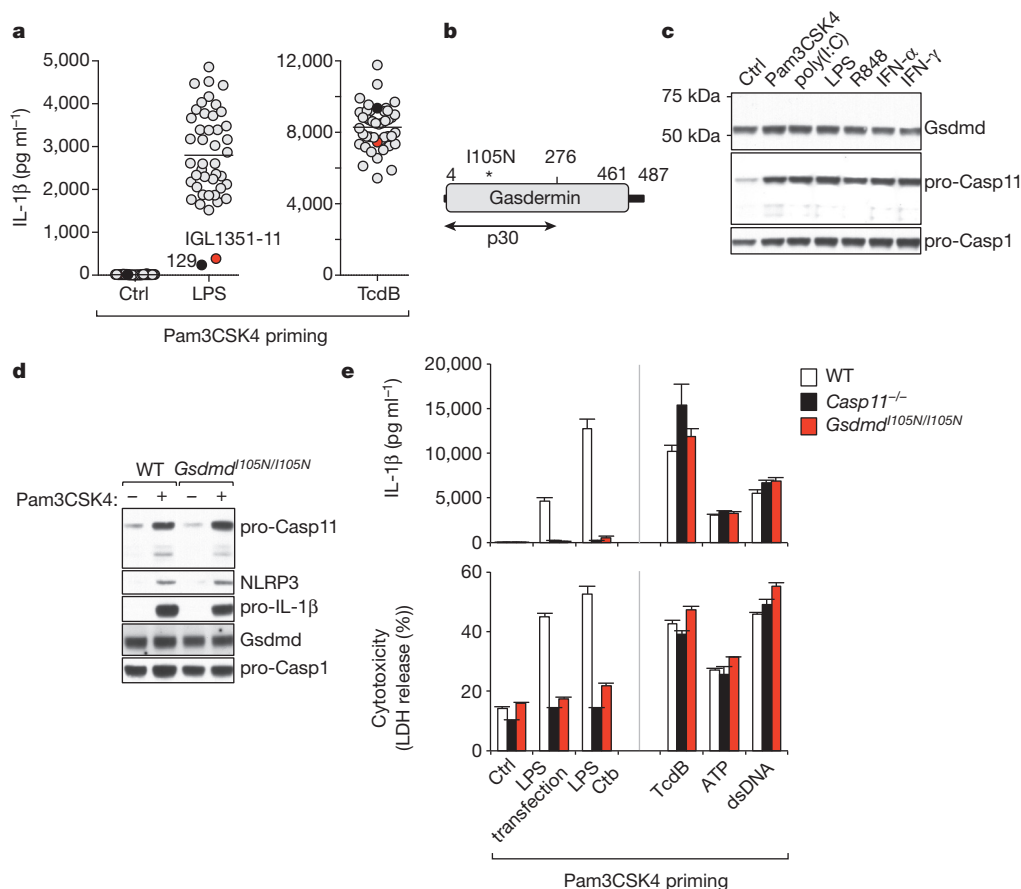


Figure 1 | *Gsdmd* mutation I105N abolishes non-canonical inflammasome signalling. **a**, Screening of third-generation offspring from ENU-treated C57BL/6 mice. Graphs indicate IL-1 β released from peritoneal macrophages cultured for 16 h. Circles indicate *Casp11* mutant 129 (black), IGL1351-11 (red), and other G3 offspring (grey). Centre bars represent averages. **b**, Schematic of mouse Gsdmd. Gasdermin domain is grey (Pfam PF04598).

Gsdmd is essential for caspase-11 signalling

We confirmed a critical role for Gsdmd in non-canonical inflammasome signalling using gene-targeted *Gsdmd*^{-/-} mice. In control experiments, Pam3CSK4-primed wild-type, *Gsdmd*^{-/-}, and *Casp11*^{-/-} BMDMs exhibited equivalent caspase-1-dependent IL-1 β secretion and pyroptosis after a 16 h treatment with TcdB, ATP, dsDNA, or the NLRC4 inflammasome trigger flagellin¹⁸ (Fig. 2a and Extended Data Fig. 3a, b). The effect of Gsdmd deficiency on caspase-1-dependent pyroptosis at earlier time points is discussed in a later section. Caspase-11 or Gsdmd deficiency also had no effect on IL-1 β secretion in response to other canonical NLRP3 inflammasome stimuli including listeriolysin O (LLO) toxin, monosodium urate (MSU), lysomotropic agent HLLoMe, or calcium pyrophosphate (CPPD)^{1,19–22} (Extended Data Fig. 3c, d). In keeping with these data, TLR agonists induced NLRP3, pro-caspase-11, and RANTES to a similar extent in wild-type and *Gsdmd*^{-/-} BMDMs (Extended Data Fig. 3a, e). Therefore, Gsdmd is dispensable for normal TLR priming. *Gsdmd*^{-/-} BMDMs were, however, resistant to pyroptosis and failed to secrete IL-1 β when transfected with *E. coli* LPS or synthetic monophosphoryl lipid A, the active moiety of LPS^{7,8} (Fig. 2a, Extended Data Fig. 3b). *Gsdmd*^{-/-} BMDMs were also unresponsive to electroporated LPS, or LPS transfected with cholera toxin B complex⁷, or *S. typhimurium* LPS (Extended Data Fig. 3f). *Gsdmd*, like *Casp11*, was also required for processing of pro-caspase-1 and pro-IL-1 β in response to cytoplasmic LPS (Fig. 2b). Finally, *Gsdmd*^{-/-} BMDMs further resembled *Casp11*^{-/-} BMDMs in failing to secrete IL-1 β or undergo pyroptosis upon infection with Gram-negative bacteria (*E. coli*, *Citrobacter rodentium*, or *Shigella flexneri*), whereas both genotypes

c, d, Immunoblots of wild-type (WT) BMDMs (**c**) or the genotypes indicated (**d**) after 6 h culture. **e**, IL-1 β and LDH released from BMDMs after 16 h. Graphs show mean \pm s.d. of triplicate wells and represent three independent experiments. LPS Ctb, LPS plus cholera toxin B complex. For source gels of **c** and **d**, see Supplementary Fig. 1.

responded normally to *Pseudomonas aeruginosa* that activates the NLRC4 canonical inflammasome^{1,3,7,23} (Fig. 2c). Collectively, these data confirm that macrophages require Gsdmd for pyroptosis and IL-1 β secretion in response to intracellular LPS.

Human caspase-4 and caspase-5, the orthologues of mouse caspase-11, induce pyroptosis in myeloid and endothelial cells in response to intracellular LPS^{9,24}. Like caspase-4/5/11, *Gsdmd* is specific to mammals^{14,25}, suggesting a co-emergence of those genes. Human GSDMD exhibits 58% amino acid sequence identity with mouse Gsdmd. Deletion of *GSDMD* or *CASP4* by CRISPR/Cas9 in human EA.hy926 endothelial cells or human THP-1 monocytes did not affect death induced by APO2L²⁶ or dsDNA, but it completely abolished intracellular LPS-induced cytotoxicity (Fig. 2d, Extended Data Fig. 3g–i). Thus, human GSDMD is required for the caspase-4-dependent response to cytoplasmic LPS.

Caspase-11 cleaves and activates Gsdmd

We next questioned how Gsdmd promotes pyroptosis and caspase-1 activation by the non-canonical inflammasome. We noticed that BMDMs transfected with LPS exhibited caspase-11-dependent cleavage of Gsdmd, giving rise to a 30-kDa N-terminal Gsdmd fragment (Fig. 3a). These data identified Gsdmd as a potential substrate of caspase-11. Recombinant caspase-11 cleaves artificial peptide substrates after an aspartate residue^{27,28}, but physiological substrates of caspase-11 are not well defined. Purified recombinant Gsdmd was cleaved by recombinant caspase-11 *in vitro* into two fragments. Edman sequencing mapped the cleavage site in Gsdmd to LLSD₂₇₆↓G₂₇₇ (Fig. 3b). The

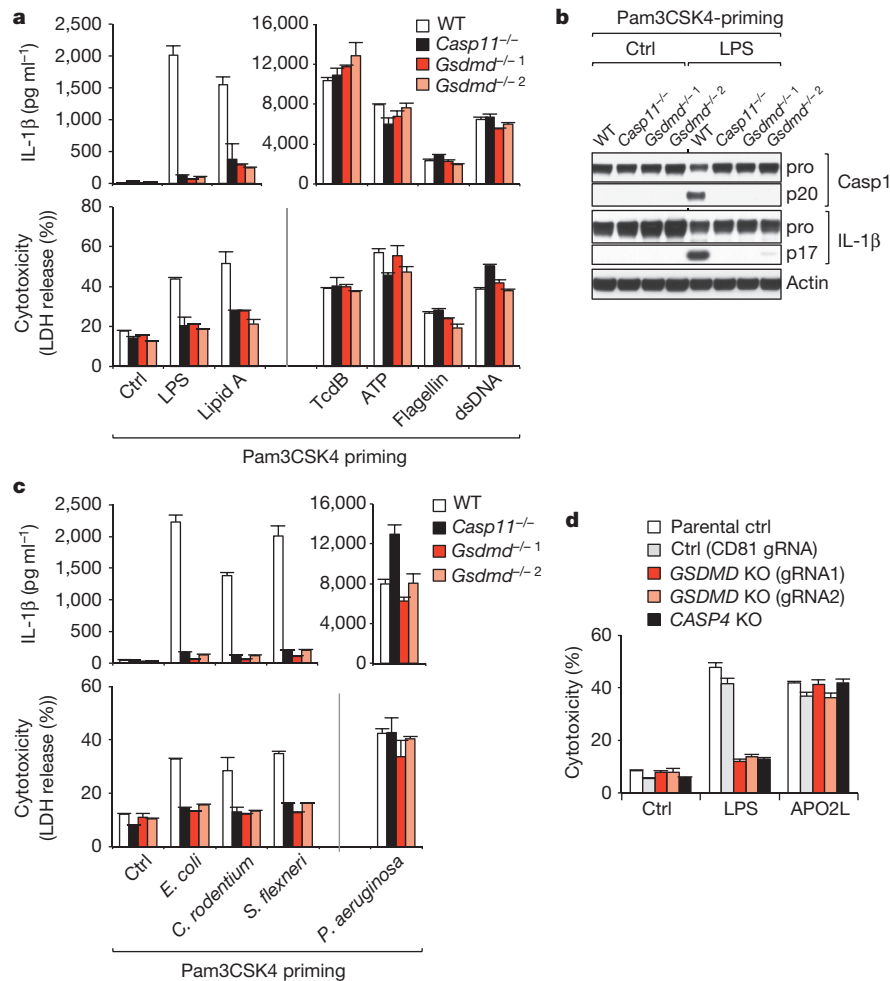


Figure 2 | Gsdmd is essential for non-canonical inflammasome signalling.

a, c, IL-1 β and LDH released from BMDMs after 16 h. *P. aeruginosa* infection was analysed after 4 h. *Gsdmd*^{-/-1} and *Gsdmd*^{-/-2} are independent knockout strains. **b**, Immunoblots of pooled cell extracts and supernatants from BMDMs at 6 h after electroporation with LPS. For source gels, see Supplementary Fig. 1.

d, Cytotoxicity of APO2L or transfected LPS on control, *GSDMD*-knockout, or *CASP4*-knockout EA.hy926 cells after 16 h. Graphs show mean \pm s.d. of triplicate wells and represent three independent experiments. gRNA, single-guide RNA.

P₁-Asp cleavage residue and P₁'-Gly residues are conserved across species (Extended Data Fig. 4a). Ectopic expression of the N-terminal Gsdmd cleavage fragment (mouse residues 1–276) killed HEK293T cells, whereas expression of full-length Gsdmd or the C-terminal cleavage fragment did not (Fig. 3c). Collectively, these data indicate that caspase-11 directly cleaves the 53-kDa inactive precursor form of Gsdmd (pro-Gsdmd) to generate the pro-pyptotic N-terminal fragment (Gsdmd p30, Fig. 1b). In full-length Gsdmd, the C terminus may mask key sites in the N terminus until cleavage at D₂₇₆G₂₇₇ releases this inhibition. Consistent with this notion, reconstitution of immortalized *Gsdmd*^{-/-} macrophages with wild-type Gsdmd restored LPS responsiveness, as evidenced by lactate dehydrogenase (LDH) release, IL-1 β secretion and caspase-1 processing, whereas reconstitution with the Gsdmd processing mutant D276A did not (Fig. 3d, e). In addition, mutation of Gsdmd within the p30 fragment (I105N) attenuated pyroptosis and caspase-1/IL-1 β processing (Fig. 1b, e and Extended Data Fig. 2b) without affecting Gsdmd cleavage (Extended Data Fig. 4b). These data confirm that conversion of pro-Gsdmd to its mature p30 form by caspase-11 is essential for both pyroptosis and caspase-1 activation in response to cytoplasmic LPS.

Cytoplasmic LPS also caused caspase-4-dependent processing of GSDMD in human THP-1 monocytes (Extended Data Fig. 5a). Furthermore, transient overexpression of caspase-4 or caspase-5 caused

human GSDMD cleavage to the pro-pyptotic p30 fragment (Extended Data Fig. 5b and c). These data together with the conserved P₁ and P₁' residues in GSDMD (Extended Data Fig. 4a) indicate a conserved mechanism of GSDMD activation by caspase-4/5/11.

Cytoplasmic LPS induces caspase-11-dependent pyroptosis and secretion of IL-1 β , but only the latter requires caspase-1, NLRP3 and ASC^{1,7}. Loss of NLRP3, ASC, or caspase-1 in BMDMs did not markedly affect caspase-11-dependent generation of mature Gsdmd p30 in response to cytoplasmic LPS (Fig. 4a), indicating that Gsdmd cleavage occurs upstream of NLRP3 inflammasome activation. The question arises as to whether Gsdmd induces pyroptosis and NLRP3 inflammasome activation simultaneously in the same cell, or if Gsdmd-mediated pyroptosis releases damage-associated molecular patterns²⁹ that trigger caspase-11-independent NLRP3 inflammasome activation in neighbouring cells (Extended Data Fig. 6a). We believe the latter hypothesis is unlikely because LPS-stimulated *Il1b*^{-/-} BMDMs exhibited normal pyroptosis, yet their culture supernatant failed to induce IL-1 β secretion from *Casp11*^{-/-} BMDMs (Extended Data Fig. 6b, c). In addition, *Casp11*^{-/-} BMDMs co-cultured with *Il1b*^{-/-} BMDMs did not secrete IL-1 β in response to cytoplasmic LPS (Fig. 4b and Extended Data Fig. 6d), making it unlikely that a paracrine signalling mechanism activates the NLRP3 inflammasome³⁰. Collectively, our data demonstrate that Gsdmd maturation by caspase-11 triggers two distinct cell-intrinsic signals:

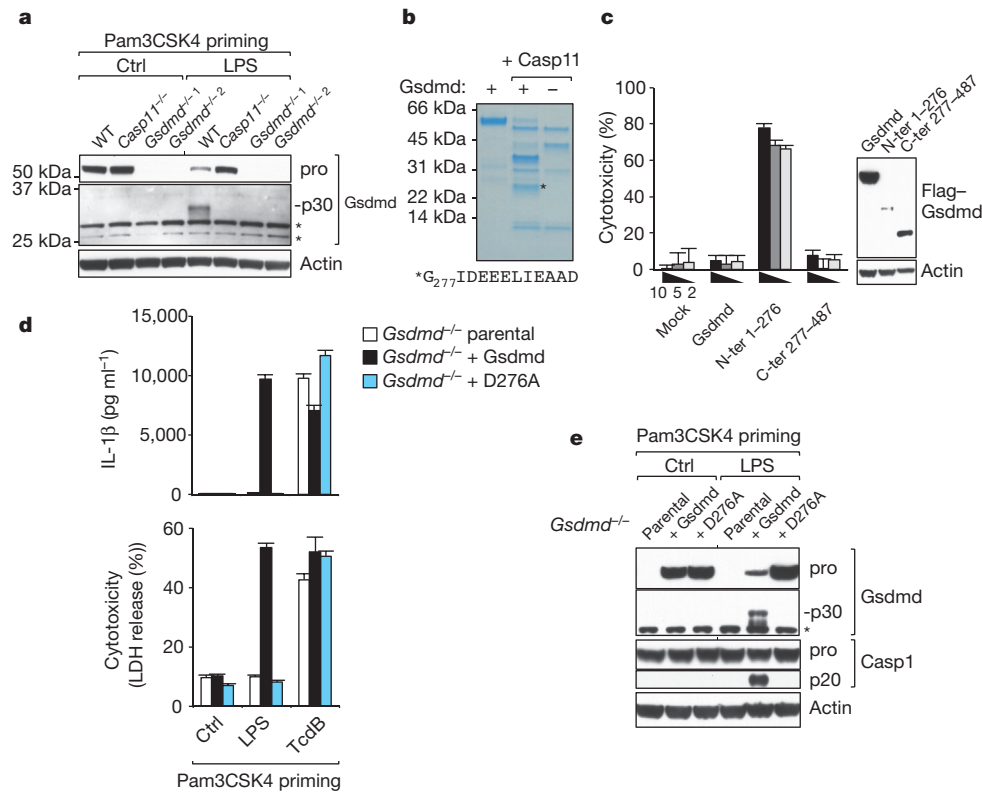


Figure 3 | Caspase-11 cleaves Gsdmd. **a**, Immunoblots of BMDM extracts and supernatants at 6 h after LPS electroporation. Non-specific bands are indicated with asterisks. **b**, Coomassie blue staining of Gsdmd incubated with caspase-11. N terminus of the band marked with an asterisk is shown. **c**, Cytotoxicity of mouse Gsdmd in HEK293Ts. Numbers indicate nanograms of plasmid transfected. **d**, IL-1β and LDH released from reconstituted

Gsdmd^{-/-} immortalized macrophages at 16 h after LPS transfection. **e**, Immunoblots of immortalized macrophage extracts and supernatants in **d**. Non-specific band is indicated with an asterisk. Graphs show mean ± s.d. of triplicate wells and represent three independent experiments. For source gels of **a**, **c**, and **e**, see Supplementary Fig. 1.

(i) pyroptosis induction; and (ii) NLRP3-dependent caspase-1 activation (Fig. 4c).

Role for Gsdmd in canonical pyroptosis

Human caspase-1 is reported to cleave GSDMD at the same P₁-Asp residue³¹ as caspase-11, albeit that the preferred tetrapeptide cleavage sequence of caspase-1 differs from that of caspase-11^{27,28}. Consistent with this study³¹, the Gsdmd p30 fragment appeared in BMDMs treated with canonical inflammasome activators including ATP (Fig. 5a and Extended Data Fig. 7a). Although *Gsdmd*^{-/-} BMDMs stimulated with ATP or flagellin underwent pyroptosis normally after 8 h, they exhibited less pyroptosis than wild-type BMDMs at earlier

time points (Figs 2a and 5c and Extended Data Fig. 7b). These data suggest that Gsdmd is also a physiological substrate of caspase-1. However, given that caspase-1-dependent pyroptosis is delayed rather than prevented, we believe that other ill-defined caspase-1 substrates can also mediate pyroptosis (Fig. 5d). This notion is in keeping with the fact that caspase-1 orthologues are found in vertebrates including non-mammals, whereas caspase-4/5/11 and *Gsdmd* are exclusive to mammals^{14,25,32}. Consistent with IL-1β being a direct substrate of caspase-1³³, *Gsdmd* deficiency did not impact caspase-1-dependent IL-1β processing in response to ATP (Fig. 5b). However, *Gsdmd*^{-/-} BMDMs secreted less IL-1β than wild-type at the 2-h time points in response to ATP and flagellin (Extended Data Fig. 7c). Gsdmd may

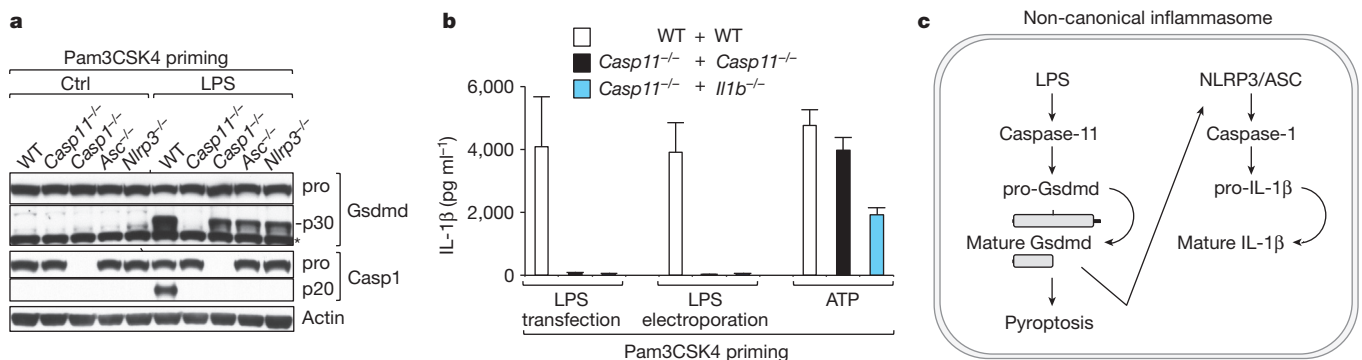


Figure 4 | Cell-intrinsic NLRP3 inflammasome activation. **a**, Immunoblots of BMDM extracts and supernatants at 6 h after electroporation with LPS. Non-specific band is indicated with an asterisk. For source gels, see Supplementary Fig. 1. **b**, IL-1β released from 1:1 mixed BMDM cultures at 16 h after

stimulation. **c**, Model for non-canonical inflammasome signalling. Graph shows mean ± s.d. of triplicate wells and represents three independent experiments.

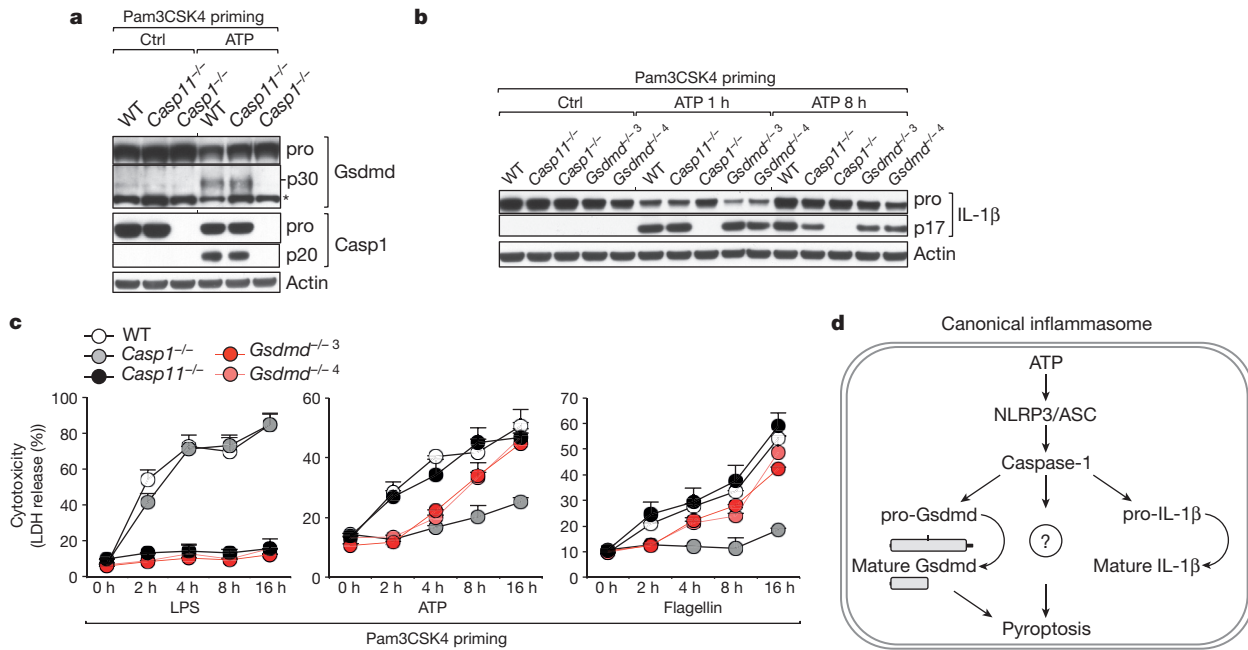


Figure 5 | Role of Gsdmd in canonical inflammasome signalling.

a, b, Immunoblots of BMDM extracts and supernatants after stimulation with ATP for 8 h (**a**) or as indicated (**b**). Non-specific band is indicated with an asterisk. For source gels, see Supplementary Fig. 2. **c**, LDH released from

BMDMs after LPS electroporation or the treatments indicated. *Gsdmd*^{-/-3} and *Gsdmd*^{-/-4} are independent knockout strains. **d**, Model for canonical inflammasome signalling. Graphs show mean ± s.d. of triplicate wells and represent three independent experiments.

contribute to IL-1β release by damaging cell membranes or regulating the ill-characterized IL-1β secretion mechanism^{34,35}.

Gsdmd is essential for lethal sepsis

We investigated the relevance of our *in vitro* findings in a mouse model of acute septic shock wherein caspase-11-dependent pyroptosis leads to lethal endotoxemia^{1,36}. Mice lacking either *Casp11* or *Gsdmd* were resistant to LPS-induced lethal septic shock (Fig. 6 and Extended Data Table 2). This result is consistent with Gsdmd being a critical pro-pyrototic substrate of caspase-11 *in vivo*.

Discussion

Pyroptosis plays an important role in anti-bacterial innate immune defence and lethal endotoxemia^{1,37,38}, but how inflammatory caspases 1, 4, 5 and 11 cause pyroptosis has remained unknown. Our data reveal that proteolytic cleavage of Gsdmd by mouse caspase-11 or human caspase-4 is essential for pyroptosis of innate immune cells and endothelial cells harbouring LPS-tainted cytoplasm. Cleaved Gsdmd also triggers NLRP3-dependent activation of caspase-1 through a cell-intrinsic pathway (Fig. 4c), although the exact mechanism responsible for NLRP3

activation remains unclear. NLRP3 senses diverse stimuli that perturb intracellular homeostasis^{6,19,21,22}, so its activation may be an indirect consequence of the effects of the Gsdmd p30 fragment. Future studies will need to address exactly what the p30 fragment does in cells. The re-defined non-canonical inflammasome signalling pathway (Fig. 4c) refutes the previous hypothesis that a caspase-1/11 complex is responsible for caspase-1 activation^{1,36}.

The partial contribution of Gsdmd to caspase-1-dependent pyroptosis suggests the existence of other pro-pyrototic caspase-1 substrates (Fig. 5d). Caspase-1 can trigger pyroptosis in the absence of Gsdmd, which is not unexpected because non-mammalian vertebrates lack *Gsdmd* yet still exhibit caspase-1-dependent pyroptosis^{14,25,32,39}. The co-emergence of caspase-4/5/11 and *Gsdmd* genes in the early stages of mammalian evolution might indicate co-evolution, perhaps in response to the need for additional anti-pathogen responses against Gram-negative bacteria.

Like other caspases, caspase-11 cleaves multiple substrates^{28,31,40}. However, our *in vitro* and *in vivo* data show that caspase-11 relies exclusively on Gsdmd to promote pyroptosis, caspase-1 activation and LPS-induced lethal sepsis. Thus Gsdmd is revealed as an unexpected, but critical facet of the anti-bacterial response.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 14 July; accepted 4 September 2015.

Published online 16 September 2015.

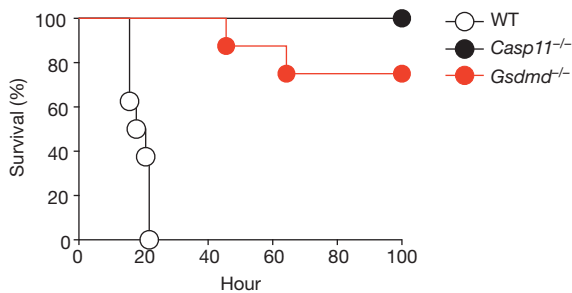


Figure 6 | Gsdmd deficiency protects against lethal sepsis. Kaplan-Meier survival plots for mice ($n = 8$ for each genotype) challenged with 54 mg kg^{-1} LPS. *Gsdmd*^{-/-6-13} mice were used (described in Extended Data Fig. 8). Statistical analysis and P values (Extended Data Table 2) were adjusted to account for multiple comparisons using Bonferroni's correction.

- Kayagaki, N. *et al.* Non-canonical inflammasome activation targets caspase-11. *Nature* **479**, 117–121 (2011).
- Broz, P. *et al.* Caspase-11 increases susceptibility to *Salmonella* infection in the absence of caspase-1. *Nature* **490**, 288–291 (2012).
- Rathinam, V. A. *et al.* TRIF licenses caspase-11-dependent NLRP3 inflammasome activation by gram-negative bacteria. *Cell* **150**, 606–619 (2012).
- Aachoui, Y. *et al.* Caspase-11 protects against bacteria that escape the vacuole. *Science* **339**, 975–978 (2013).
- Mariathasan, S. *et al.* Differential activation of the inflammasome by caspase-1 adaptors ASC and Ipaf. *Nature* **430**, 213–218 (2004).
- Mariathasan, S. *et al.* Cryopyrin activates the inflammasome in response to toxins and ATP. *Nature* **440**, 228–232 (2006).

7. Kayagaki, N. *et al.* Noncanonical inflammasome activation by intracellular LPS independent of TLR4. *Science* **341**, 1246–1249 (2013).
8. Hagar, J. A., Powell, D. A., Achoui, Y., Ernst, R. K. & Miao, E. A. Cytoplasmic LPS activates caspase-11: implications in TLR4-independent endotoxic shock. *Science* **341**, 1250–1253 (2013).
9. Shi, J. *et al.* Inflammatory caspases are innate immune receptors for intracellular LPS. *Nature* **514**, 187–192 (2014).
10. Nelms, K. A. & Goodnow, C. C. Genome-wide ENU mutagenesis to reveal immune regulators. *Immunity* **15**, 409–418 (2001).
11. Xu, H. *et al.* Innate immune sensing of bacterial modifications of Rho GTPases by the PIR1 inflammasome. *Nature* **513**, 237–241 (2014).
12. Katoh, M. & Katoh, M. Identification and characterization of human DFNA5L, mouse Dfna5l, and rat Dfna5l genes *in silico*. *Int. J. Oncol.* **25**, 765–770 (2004).
13. Fujii, T. *et al.* Gasdermin D (Gsdmd) is dispensable for mouse intestinal epithelium development. *Genesis* **46**, 418–423 (2008).
14. Tamura, M. *et al.* Members of a novel gene family, *Gsdm*, are expressed exclusively in the epithelium of the skin and gastrointestinal tract in a highly tissue-specific manner. *Genomics* **89**, 618–629 (2007).
15. Roberts, T. L. *et al.* HIN-200 proteins regulate caspase activation in response to foreign cytoplasmic DNA. *Science* **323**, 1057–1060 (2009).
16. Fernandes-Alnemri, T. *et al.* The AIM2 inflammasome is critical for innate immunity to *Francisella tularensis*. *Nature Immunol.* **11**, 385–393 (2010).
17. Hornung, V. *et al.* AIM2 recognizes cytosolic dsDNA and forms a caspase-1-activating inflammasome with ASC. *Nature* **458**, 514–518 (2009).
18. Franchi, L. *et al.* Cytosolic flagellin requires Ipaf for activation of caspase-1 and interleukin 1 β in salmonella-infected macrophages. *Nature Immunol.* **7**, 576–582 (2006).
19. Meixenberger, K. *et al.* *Listeria monocytogenes*-infected human peripheral blood mononuclear cells produce IL-1 β , depending on listeriolysin O and NLRP3. *J. Immunol.* **184**, 922–930 (2010).
20. Schroder, K. & Tschopp, J. The inflammasomes. *Cell* **140**, 821–832 (2010).
21. Martinon, F., Petrilli, V., Mayor, A., Tardivel, A. & Tschopp, J. Gout-associated uric acid crystals activate the NALP3 inflammasome. *Nature* **440**, 237–241 (2006).
22. Hornung, V. *et al.* Silica crystals and aluminum salts activate the NALP3 inflammasome through phagosomal destabilization. *Nature Immunol.* **9**, 847–856 (2008).
23. Sutterwala, F. S. *et al.* Immune recognition of *Pseudomonas aeruginosa* mediated by the IPAF/NLRC4 inflammasome. *J. Exp. Med.* **204**, 3235–3245 (2007).
24. Casson, C. N. *et al.* Human caspase-4 mediates noncanonical inflammasome activation against gram-negative bacterial pathogens. *Proc. Natl Acad. Sci. USA* **112**, 6688–6693 (2015).
25. Sakamaki, K. & Satou, Y. Caspases: evolutionary aspects of their functions in vertebrates. *J. Fish Biol.* **74**, 727–753 (2009).
26. Ashkenazi, A. & Dixit, V. M. Death receptors: signaling and modulation. *Science* **281**, 1305–1308 (1998).
27. Thornberry, N. A. *et al.* A combinatorial approach defines specificities of members of the caspase family and granzyme B. Functional relationships established for key mediators of apoptosis. *J. Biol. Chem.* **272**, 17907–17911 (1997).
28. Kang, S. J. *et al.* Dual role of caspase-11 in mediating activation of caspase-1 and caspase-3 under pathological conditions. *J. Cell Biol.* **149**, 613–622 (2000).
29. Kaczmarek, A., Vandenabeele, P. & Krysko, D. V. Necroptosis: the release of damage-associated molecular patterns and its physiological relevance. *Immunity* **38**, 209–223 (2013).
30. Rühl, S. & Broz, P. Caspase-11 activates a canonical NLRP3 inflammasome by promoting K⁺ efflux. *Eur. J. Immunol.* <http://dx.doi.org/10.1002/eji.201545772> (2015).
31. Agard, N. J., Maltby, D. & Wells, J. A. Inflammatory stimuli regulate caspase substrate profiles. *Mol. Cell. Proteomics* **9**, 880–893 (2010).
32. Angosto, D. *et al.* Evolution of inflammasome functions in vertebrates: Inflammasome and caspase-1 trigger fish macrophage cell death but are dispensable for the processing of IL-1 β . *Innate Immun.* **18**, 815–824 (2012).
33. Thornberry, N. A. *et al.* A novel heterodimeric cysteine protease is required for interleukin-1 β processing in monocytes. *Nature* **356**, 768–774 (1992).
34. Eder, C. Mechanisms of interleukin-1 β release. *Immunobiology* **214**, 543–553 (2009).
35. Liu, T. *et al.* Single-cell imaging of caspase-1 dynamics reveals an all-or-none inflammasome signaling response. *Cell Rep.* **8**, 974–982 (2014).
36. Wang, S. *et al.* Murine caspase-11, an ICE-interacting protease, is essential for the activation of ICE. *Cell* **92**, 501–509 (1998).
37. Miao, E. A. *et al.* Caspase-1-induced pyroptosis is an innate immune effector mechanism against intracellular bacteria. *Nature Immunol.* **11**, 1136–1142 (2010).
38. Bergsbaken, T., Fink, S. L. & Cookson, B. T. Pyroptosis: host cell death and inflammation. *Nature Rev. Microbiol.* **7**, 99–109 (2009).
39. Xie, H. X. *et al.* *Edwardsiella tarda*-induced cytotoxicity depends on its type III secretion system and flagellin. *Infect. Immun.* **82**, 3436–3445 (2014).
40. Py, B. F. *et al.* Caspase-11 controls interleukin-1 β release through degradation of TRP1. *Cell Rep.* **6**, 1122–1128 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank the staff of the Australian Phenomics Facility, Genentech Transgenic Technology and FACS cores, and K. Bowman, J. Payandeh, E. Dueber, R. Aglietti, A. Gupta and A. Peterson for technical expertise and discussion, K. Newton for manuscript editing, A. Muszyński, L. S. Forsberg, and R. W. Carlson for *S. typhimurium* LPS. Most authors were employees of Genentech, Inc.

Author Contributions N.K., I.B.S., B.L.L., K.O., T.C., B.H., P.S.L., Q.T.P., J.R.L., H.L., J.W., S.K., J.Z., W.P.L., S.J.S., L.X.M., L.F., Y.Z. and E.M.B. designed and performed experiments. K.A. and S.W. generated *Gsdmd*^{−/−} mice. S.W. and M.R.-G. generated Casp1^{−/−} mice. N.K. and E.M.B. prepared the manuscript. N.K., G.S.S., E.M.B., C.C.G. and V.M.D. contributed to the study design and data analyses.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to N.K. (kayagaki@gene.com) or V.M.D. (dixit@gene.com).

METHODS

ENU-mutagenized mouse strains. Third-generation (G3) offspring used for the phenotypic screen were from ENU-treated C57BL/6 mice as described¹⁰. 129X1/SvJ strain was used as a *Casp11* mutant¹ control. All animals were housed under specific-pathogen-free conditions at the Australian Phenomics Facility. Animal experiments were approved by the Australian National University Animal Ethics and Experimentation Committee.

Exome capture and sequencing of G1 founder. Exome-enriched, paired-end libraries were prepared from genomic DNA of first generation (G1) mice using the Agilent SureSelect XT2 Mouse All Exon kit (Agilent) following the manufacturer's instructions. Each sample was prepared with an index and then pooled in a batch of six in equimolar amounts before exome enrichment. Each 6-plex exome-enriched library was sequenced in one lane of an Illumina HiSeq2000 with version 2 chemistry as 100-bp paired-end reads. ENU variants were identified as described previously⁴¹. In brief, sequence reads were mapped to the GRCm38 assembly of the reference mouse genome using the default parameters of the Burrows-Wheeler Aligner (BWA, <http://bio-bwa.sourceforge.net>). Untrimmed reads were aligned allowing a maximum of two sequence mismatches and were discarded where they aligned to the genome more than once. Sequence variants were identified with SAMtools (<http://samtools.sourceforge.net>) and annotated using Annovar (<http://www.openbioinformatics.org>). PolyPhen2 (<http://genetics.bwh.harvard.edu/pph2>) and SIFT (<http://sift.jcvi.org/>) were used for the calculation of variant effect. Variants were filtered to prioritize novel variants not in dbSNP (or in a list of common mouse variants identified by the pipeline) and predicted to be deleterious to the protein by PolyPhen and/or SIFT. For variant validation, G1 samples were genotyped for the ENU variants identified by the exome sequencing using a competitive, allele-specific dual FRET-based assay, KASP (LGC). For genotyping of G3 offspring and IGL1351 pedigree, primers were designed based on the SNV locus-flanking sequence. Plates were read in a FLUOstar Optima (BMG Labtechnologies) plate reader.

Other mice. *Nlrp3*^{-/-}, *Asc*^{-/-}, and *Casp11*^{-/-} mice on a C57BL/6N background were described previously¹. *Il1b*^{-/-} mice generated from AB2.2 ES cells (129S7) were obtained from Taconic and backcrossed to C57BL/6J up to the N10 generation. C57BL/6-type *Casp11* genotype was confirmed by PCR as described¹. *Gsdmd*^{-/-} mice were obtained by cytoplasmic injection of C57BL/6N zygotes with 25 ng μ l⁻¹ wild-type Cas9 mRNA (Life Technologies) and 13 ng μ l⁻¹ *in vitro*-transcribed single-guide RNA (gRNA) prepared by MEGAshortscript T7 Kit (Life Technologies). Tail DNA from resulting offspring was analysed by PCR and sequencing. Target sequences of gRNA are 5'-GAGAAGGGAAAATTCTGG (G9) and 5'-AGGG

CAGAGTGATGTTGTC (G2). Allele sequences of *Gsdmd*^{-/-} mice used are provided in Extended Data Fig. 8. *Casp1*^{-/-} mice (Extended Data Fig. 7d) lacking exons 1 and 2 were generated at Genentech from gene-targeted C57BL/6N C2 ES cells. *Casp1*^{-/-} mice were genotyped with PCR primers (5'-CCTGAATCTTAGACCAAGTTGAG; 5'-AGGCAGAAGGAATAGGAATAGT and 5'-CCAGGTATCTCAATCACATGGT) yielding a 167-bp wild-type DNA fragment and a 287-bp mutant DNA fragment. The Genentech Institutional Animal Care and Use Committee approved all animal studies.

Reagents and antibodies. Ultra-pure LPS (*E. coli* O111:B4), Pam3CSK4, poly(I:C) LMW, R837 (Imiquimod), dsDNA (poly(dA:dT)), ultra-pure flagellin (*P. aeruginosa*), MSU, and CPPD were from Invivogen. Other reagents were synthetic monophosphoryl lipid A (Enzo Life Sciences), TcdB, cholera toxin B (Ctb; List Biological Laboratories), LLO toxin (ReagentProteins), HLLome (Chem Impex International), IFN- α , (PBL Assay Science), IFN- γ (eBioscience), and ATP (Sigma). LPS from *S. typhimurium* was described previously^{42,43}. Antibodies used include: mouse caspase-1 (clone 4B4, Genentech), human caspase-1 (Bally-1, AdipoGen), caspase-4 (4B9, Enzo Life Sciences), caspase-5 (4429, Cell Signaling Technology), caspase-11 (clone 17D9, Novus Biologicals), Gsdmd (G7422 anti-Gsdmd aa 126-138, Sigma), IL-1 β (GTX74034, GeneTex), NLRP3 (clone Cryo-2, AdipoGen), and Flag epitope (M2, Sigma).

Macrophage cultures. Bone marrow cells were differentiated in DMEM supplemented with 10% endotoxin-free fetal bovine serum (Omega Scientific) and 20% M-CSF-conditioned medium for 5–6 days. Adherent BMDMs or thioglycollate-elicited peritoneal macrophages were cultured overnight in 96-well plates at 1×10^6 cells per ml before being primed for 5–6 h with 1μ g ml⁻¹ Pam3CSK4 in OPTI-MEM (Life Technologies). Primed cells were transfected with 2μ g ml⁻¹ LPS or 5μ g ml⁻¹ lipid A by using 0.25% v/v FuGENE HD (Promega) or 20μ g ml⁻¹ Ctb, or electroporated with 2μ g ml⁻¹ LPS by using the 4D-Nucleofector system (Lonza) as described previously⁷. Unless specified, FuGENE HD was used for the LPS transfection. The other conditions for inflammasome activations were 5 mM ATP, 100 ng ml⁻¹ TcdB, 2μ g ml⁻¹ dsDNA plus 0.1% v/v Lipofectamine 2000 (Life Technologies), 0.5μ g ml⁻¹ flagellin plus 0.25% v/v FuGENE HD, 1 mg ml⁻¹ MSU, 500 μ g ml⁻¹ CPPD, and 1 mM HLLome. Infections with *P. aeruginosa* (ATCC 27853; multiplicity of infection (MOI) 25), *E. coli* (ATCC 11775, MOI

30), *C. rodentium* (ATCC 51116, MOI 20), or *S. flexneri* (ATCC 9199, MOI 20) were for 1.5 h and then cultures were supplemented with 100μ g ml⁻¹ Gentamycin (Life Technologies). For stimulation with TLR agonists or interferons, BMDMs were cultured with 1μ g ml⁻¹ Pam3CSK4, 5μ g ml⁻¹ polyI:C, 5μ g ml⁻¹ LPS, 2μ g ml⁻¹ R848, 75 ng ml^{-1} IFN- α , or 500 ng ml^{-1} IFN- γ for the indicated periods. ELISAs were used to measure IL-1 β (Meso Scale Discovery) in culture supernatants. A CytoTox 96 Non-Radioactive Cytotoxicity assay (Promega) measured cell death. For measuring RANTES, Luminex assay (Cocktail (Roche)) was used. Each figure legend shows the number of samples per experiment and number of experiments that were analysed independently. For immunoblotting, cells were lysed with RIPA buffer (50 mM Tris-HCl pH 7.4, 150 mM NaCl, 1 mM EDTA, 1 \times Complete Protease Inhibitor (Roche), 1% Triton X-100, 0.1% SDS). For cleavage studies (Gsdmd p30, caspase-1 p20, and IL-1 β p17), pooled culture supernatants and cell extracts were precipitated with 7.2% trichloroacetic acid plus 0.15% sodium deoxycholate.

Recombinant proteins. DNA encoding Flag-tagged caspase-11 (residues 2–373) and His6-MBP-tagged murine Gsdmd (residues 2–487) were generated by gene synthesis, and subcloned into a modified version of pAcgp67 (BD Biosciences). Transfer vectors were co-transfected with BestBac linearized viral DNA (Expression Systems) into Sf9 cells (Life Technologies) using Transit-IT-Insect transfection reagent (Mirus Bio) to produce recombinant baculovirus. The virus was amplified twice to prepare the stock used for protein expression. Protein was expressed in *Tni* pro cells (Expression Systems). A 22 l Wave bioreactor was inoculated with *Tni* pro cells at 1×10^6 cells per ml in serum-free ESF921 (Expression Systems). At 48 h post infection, the *Tni* cells were harvested and frozen at -80°C . For the purification of Gsdmd, the baculovirus cell pellets were thawed and lysed in lysis buffer (50 mM Tris-HCl pH 8.0, 150 mM NaCl, 20 mM imidazole, 1 mM TCEP plus Protease Inhibitor Cocktail (Roche)). The cell lysate was loaded onto a 3 ml NiNTA column (QIAGEN) pre-equilibrated with buffer A (20 mM Tris pH 8.0, 300 mM NaCl, 20 mM imidazole, 0.5 mM TCEP). The column was washed with 20 column volumes buffer A, then eluted with 5 column volumes of buffer B (20 mM Tris-HCl pH 8.0, 300 mM NaCl, 250 mM imidazole, 0.5 mM TCEP). The elutant was then loaded onto a 5 ml MBP column (GE Healthcare Life Sciences) pre-equilibrated with buffer C (20 mM Tris-HCl pH 7.5, 150 mM NaCl, 10% glycerol, 0.5 mM TCEP). The MBP column was washed with 15 column volumes of buffer C. Gsdmd protein was eluted with buffer C containing 10 mM maltose. To cleave off His-MBP tag, purified His6-MBP-tagged Gsdmd was digested with TEV (Sigma). Un-cleaved Gsdmd and TEV enzyme were removed by passing through a 3 ml NiNTA column. For the purification of caspase-11, the baculovirus cell pellets were thawed and lysed in lysis buffer (50 mM Tris-HCl pH 7.5, 150 mM NaCl, 0.5 mM TCEP plus protease inhibitor cocktail) for 1 h. The cell lysate was loaded onto a 4 ml Flag column (Sigma) pre-equilibrated with buffer D (50 mM Tris-HCl pH 7.5, 150 mM NaCl, 10% glycerol, 0.5 mM TCEP). The column was washed with 5 column volumes of buffer C, then 20 column volumes of buffer E (buffer D plus 0.1% TritonX-114) followed with 15 column volumes of Buffer D. Caspase-11 was eluted with buffer F (buffer D plus 0.2 mg ml^{-1} 3 \times Flag peptide (Sigma)). The purified recombinant caspase-11 was then concentrated and further purified by a HiLoad 16/60 Superdex 200 column with buffer G (20 mM Tris-HCl pH 7.5, 150 mM NaCl, 10% glycerol). Purified recombinant caspase-11 and Gsdmd were characterized by SDS-PAGE and mass spectrometry, and stored at -20°C .

Recombinant Gsdmd processing assay. Recombinant Gsdmd (4μ M) was incubated in the presence or absence of 2μ M recombinant caspase-11 at 37°C for 30 min in standard caspase reaction buffer (20 mM PIPES pH 7.2, 10% sucrose, 100 mM NaCl, 0.1% CHAPS, 1 mM EDTA, 10 mM DTT). Samples were then subjected to SDS-page followed by Coomassie blue staining (Expediton). For Edman degradation analysis, samples were blotted onto a PVDF after SDS-page electrophoresis, and the band corresponding to the Gsdmd protein fragment was subjected to N-terminal sequencing using traditional Edman chemistries⁴⁴ on a 494HT Procise sequencer (Applied Biosystems) with fast cycles⁴⁵.

Plasmids and transient expression. cDNAs encoding Gsdmd (human and mouse), human caspase-4, and caspase-5 were artificially synthesized with a Flag epitope and subcloned into pcDNA3.1/Zeo(+) (Life Technologies) for transient expression in HEK293T cells, or into pLenti6.3/V5-DEST (Life Technologies) for lentivirus production. Human codon-optimized Cas9 was synthesized and subcloned into pLenti7.3 (Life Technologies). Mutagenesis of cDNAs was performed with a QuikChange site-directed mutagenesis kit (Agilent Technologies). For transient expression, HEK293T cells (ATCC) were cultured overnight in 96-well plates at 1.2×10^5 cells per ml, then transfected with 2–10 ng of plasmids by using 0.16 μ l Lipofectamine 2000. At 24 h after transfection, cytotoxicity against HEK293T cells was measured by CellTiter-Glo (Promega). For human GSDMD processing, HEK293T cells that stably express human GSDMD were selected using 1 mg ml^{-1} Zeocin (Invivogen) after transfection with human GSDMD plasmid. The stable

transfectant cells were transiently transfected with 20 ng of human caspase-4 or -5 plasmids as described above. At 24 h after transfection, cells were lysed with RIPA buffer for immunoblotting. HEK293T cell lines used were regularly tested for mycoplasma. All cell lines used were authenticated by short tandem repeat (STR) profiling, single nucleotide polymorphism (SNP) fingerprinting, and mycoplasma testing.

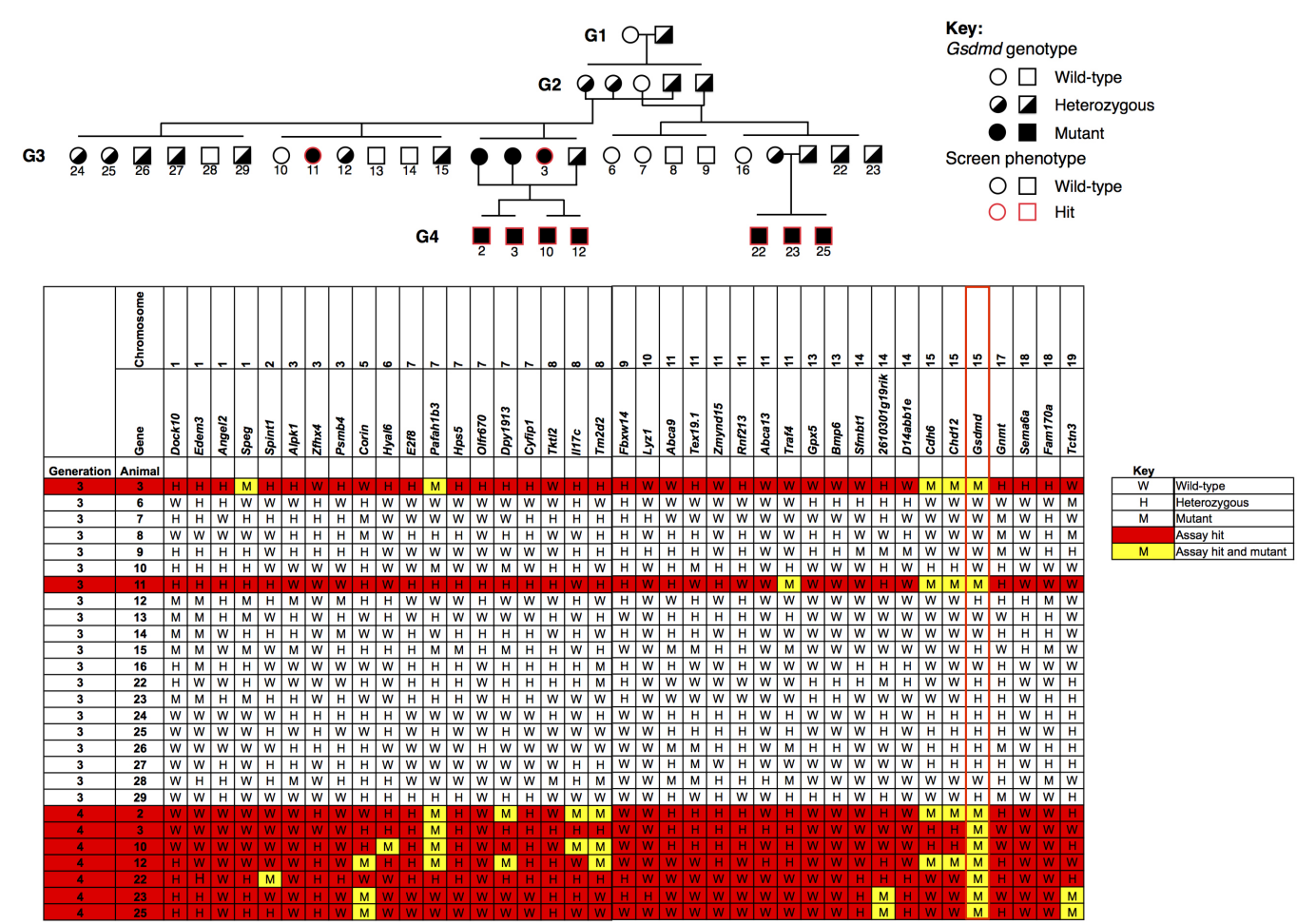
Knockout cell lines. Human EA.hy926 (ATCC) endothelial cells and human THP-1 (ATCC) monocytes were lentivirally transduced with Cas9 and single-cell cloned. gRNAs were transduced into Cas9⁺ cloned lines by lentiviral delivery with pLKO.1 vector (Sigma) followed by selection of gRNA-expressing cells by Puromycin (Life Technologies). At 11–14 days post-selection, cells were harvested for experiments. The target sequences of gRNA used are 5'-GTTGGCTTCCTGGGCTGCTA for human CD81 (control), 5'-GCATGCGAGAATCTCACGC for luciferase (control), 5'-GGTGTGTTGGATAACTTGG for *CASP4* knockout, 5'-GTAGTCCGGA GAGTGGTCC for human *GSDMD* knockout (gRNA1), and 5'-AACCACCAG GCAGTAGGGC for human *GSDMD* knockout (gRNA2). EA.hy926 cells were transfected with 10 µg ml⁻¹ LPS by 0.5% v/v Lipofectamine 2000 or treated with 100 ng ml⁻¹ APO2L (R&D systems). At 24 h after stimulation, cytotoxicity against EA.hy926 was measured by CellTiter-Glo. THP-1 cells were primed with 0.5 µg ml⁻¹ Pam3CSK4 for 4 h, then plated in 96-well plates at 1 × 10⁶ cells per ml with OPTI-MEM containing 0.5 µg ml⁻¹ Pam3CSK4, followed by transfection of 2 µg ml⁻¹ dsDNA with 0.125% v/v Lipofectamine 2000. For LPS electroporation into Pam3CSK4-primed THP-1 cells, the Neon transfection system (Life Technologies) was used with 1 × 10⁶ cells plus LPS 0.1 µg per chamber condition. All the cell lines used were regularly tested for mycoplasma. All cell lines used were authenticated by STR profiling, SNP fingerprinting, and mycoplasma testing.

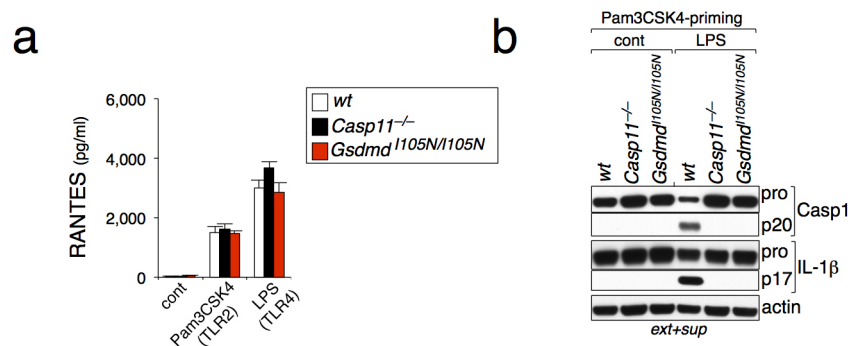
Gsdmd reconstitution. Macrophage progenitor cells from *Gsdmd*^{-/-} mice (Extended Data Fig. 8) were immortalized by ER-Hoxb8 as described previously^{46,47}. The immortalized macrophage progenitor cells were lentivirally transduced with cDNA encoding either wild-type or D276A mouse *Gsdmd* using pLenti6.3/V5-DEST vector. Cells were selected in 6.25 µg ml⁻¹ Blasticidin (Invitrogen) and

differentiated in M-CSF-containing conditioned media⁴⁸. Stimulation of ER-Hoxb8-immortalized macrophages was performed as described above. All Hoxb cell lines used were regularly tested for mycoplasma.

Endotoxic shock model. Female mice aged 8–10 weeks were injected intraperitoneally with 54 mg kg⁻¹ LPS (*E. coli* O111:B4, Sigma) and monitored eight times daily for a total of 6 days. *Gsdmd*^{-/-} mice used (*Gsdmd*^{-/-} 6–13) are described in Extended Data Fig. 8. Statistical analysis was performed with log-rank (Mantel–Cox) tests using Prism, and *P* values were adjusted to account for multiple comparisons using Bonferroni's correction. Sample sizes were chosen by standard methods to ensure adequate power and no mice were excluded from the analysis. No randomization or blinding was used for the animal studies. No statistical method was used to estimate sample size. The Genentech Institutional Animal Care and Use Committee approved all animal studies.

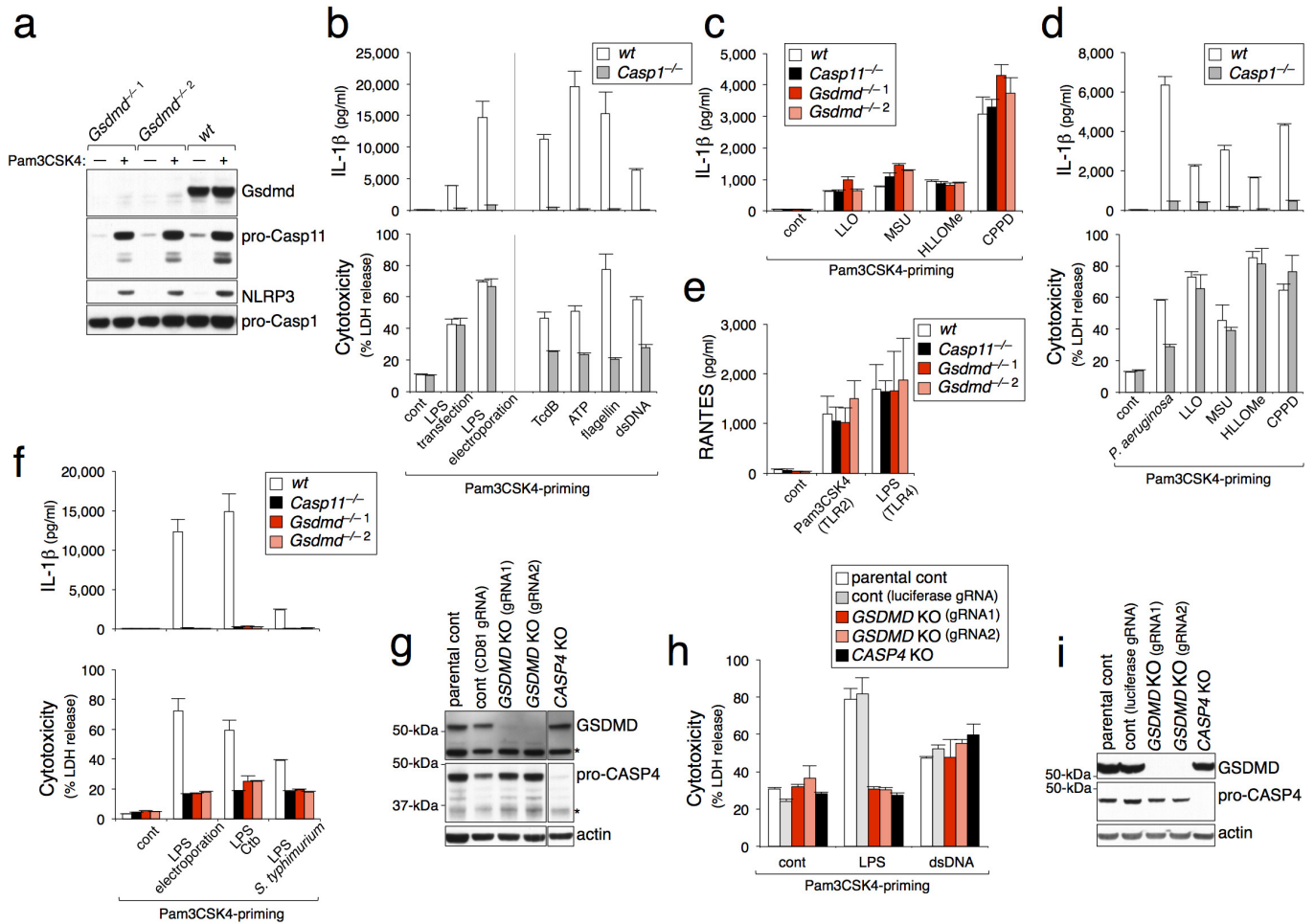
41. Andrews, T. D. *et al.* Massively parallel sequencing of the mouse exome to accurately identify rare, induced mutations: an immediate source for thousands of new mouse models. *Open Biol.* **2**, 120061 (2012).
42. Carlson, R. W., Forsberg, L. S. & Kannenberg, E. L. Lipopolysaccharides in Rhizobium-legume symbioses. *Subcell. Biochem.* **53**, 339–386 (2010).
43. Tamayo, R. *et al.* Identification of *cptA*, a *PmrA*-regulated locus required for phosphoethanolamine modification of the *Salmonella enterica* serovar typhimurium lipopolysaccharide core. *J. Bacteriol.* **187**, 3391–3399 (2005).
44. Edman, P. A method for the determination of amino acid sequence in peptides. *Arch. Biochem.* **22**, 475 (1949).
45. Henzel, W. J., Tropea, J. & Dupont, D. Protein identification using 20-minute Edman cycles and sequence mixture analysis. *Anal. Biochem.* **267**, 148–160 (1999).
46. Wang, G. G. *et al.* Quantitative production of macrophages or neutrophils *ex vivo* using conditional Hoxb8. *Nature Methods* **3**, 287–293 (2006).
47. Qu, Y. *et al.* Phosphorylation of NLRC4 is critical for inflammasome activation. *Nature* **490**, 539–542 (2012).
48. Rosas, M. *et al.* Hoxb8 conditionally immortalised macrophage lines model inflammatory monocytic cells with important similarity to dendritic cells. *Eur. J. Immunol.* **41**, 356–365 (2011).





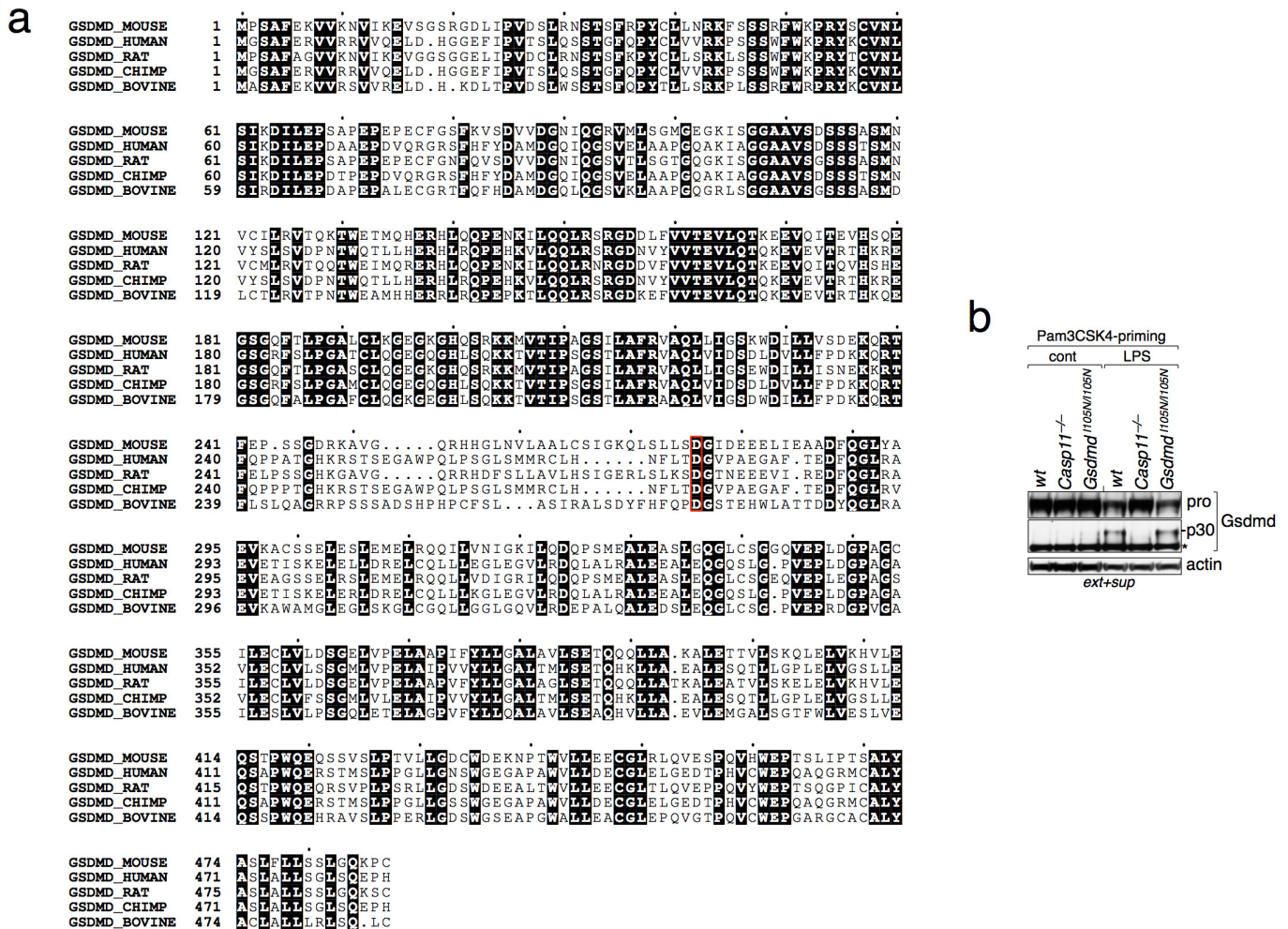
Extended Data Figure 2 | *Gsdmd*^{I105N/I105N} BMDMs respond normally to TLR agonists. **a**, RANTES production from BMDMs cultured for 16 h with medium alone (cont) or the TLR stimulants indicated. **b**, Western blots of

BMDM extracts and supernatants at 6 h after LPS electroporation. Graph shows the mean \pm s.d. of triplicate wells and represents three independent experiments. For source gel, see Supplementary Fig. 2.



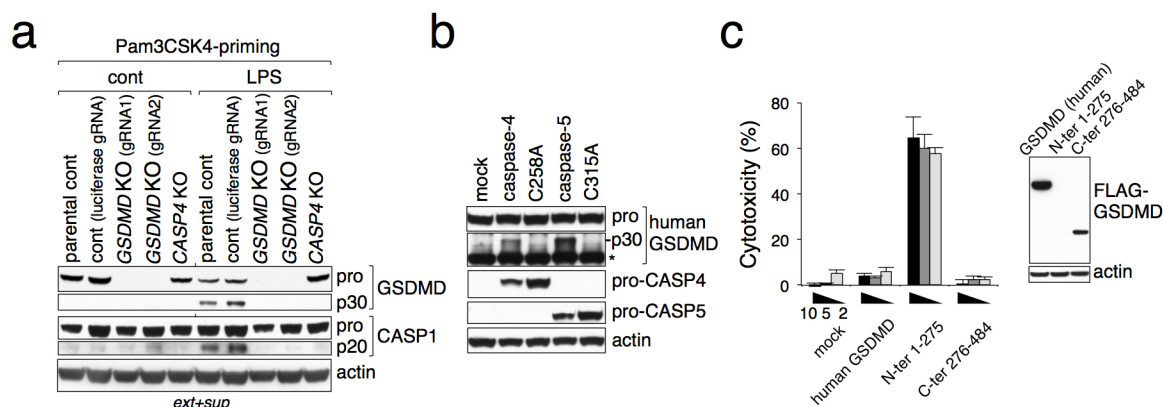
Extended Data Figure 3 | Non-canonical inflammasome signalling requires *Gsdmd*. **a**, Western blots of BMDMs cultured with or without Pam3CSK4 for 6 h. *Gsdmd*^{-/-1} and *Gsdmd*^{-/-2} are independent knockout strains. **b–d**, IL-1 β and LDH released from BMDMs after 16 h. *P. aeruginosa* infection was analysed at 4 h. **e**, RANTES production from BMDMs after 16 h. **f**, IL-1 β and LDH released from BMDMs at 16 h after LPS electroporation, LPS plus cholera toxin B (Ctb) complex, or *S. typhimurium* LPS transfection.

g, i, Western blots of EA.hy926 (**g**) or THP-1 cells (**i**). **h**, LDH released from control (parental or luciferase gRNA), *GSDMD* knockout (KO), or *CASP4* knockout THP-1 cells at 16 h after LPS electroporation or dsDNA transfection. Graphs show mean \pm s.d. of triplicate wells and represent three independent experiments. For source gels of **a**, **g** and **i**, see Supplementary Information Fig. 2.



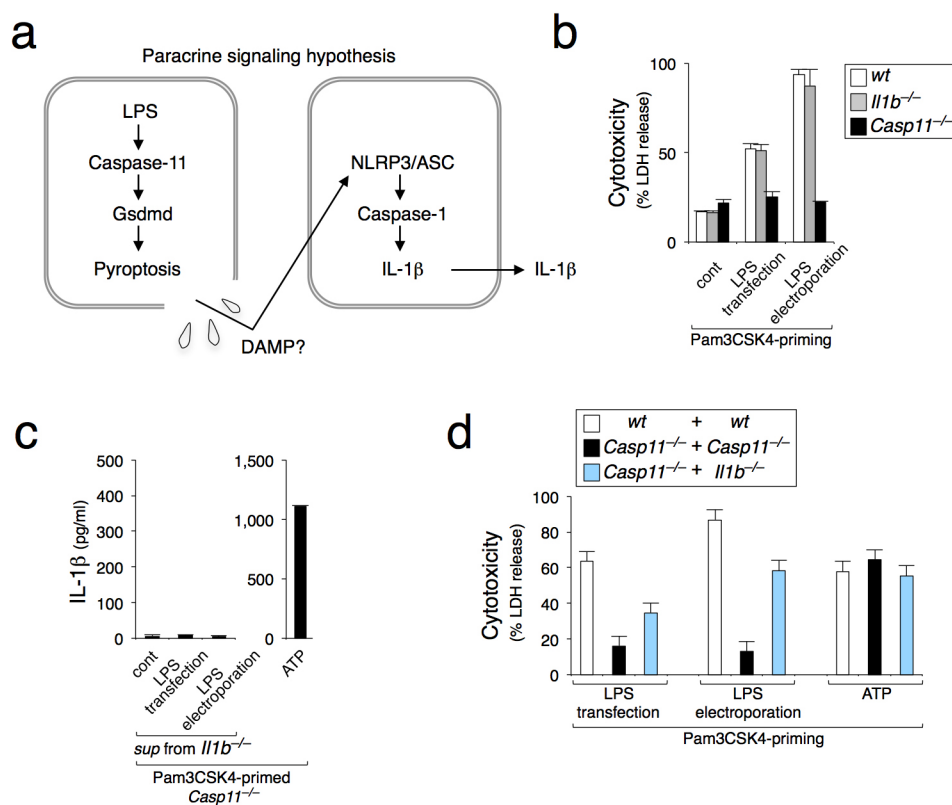
Extended Data Figure 4 | Gsdmd cleavage site. **a**, Alignment of Gsdmd amino acid sequences. The conserved Gsdmd cleavage site (D276 in mouse) is boxed in red. **b**, Immunoblots of BMDM extracts and supernatants at 16 h after

LPS/Ctb treatment. A non-specific band is indicated with an asterisk. Cont, Ctb alone. For source gel, see Supplementary Fig. 3.



Extended Data Figure 5 | Human GSDMD processing. **a**, Western blots of THP-1 extracts and supernatants at 3 h after LPS electroporation. **b**, Western blots of human GSDMD/HEK293T stable transfectants at 24 h after transient transfection of the indicated plasmids. A non-specific band is indicated with an asterisk. **c**, Cytotoxicity of the human proteins indicated at 24 h after transient transfection of HEK293T cells. Numbers indicate

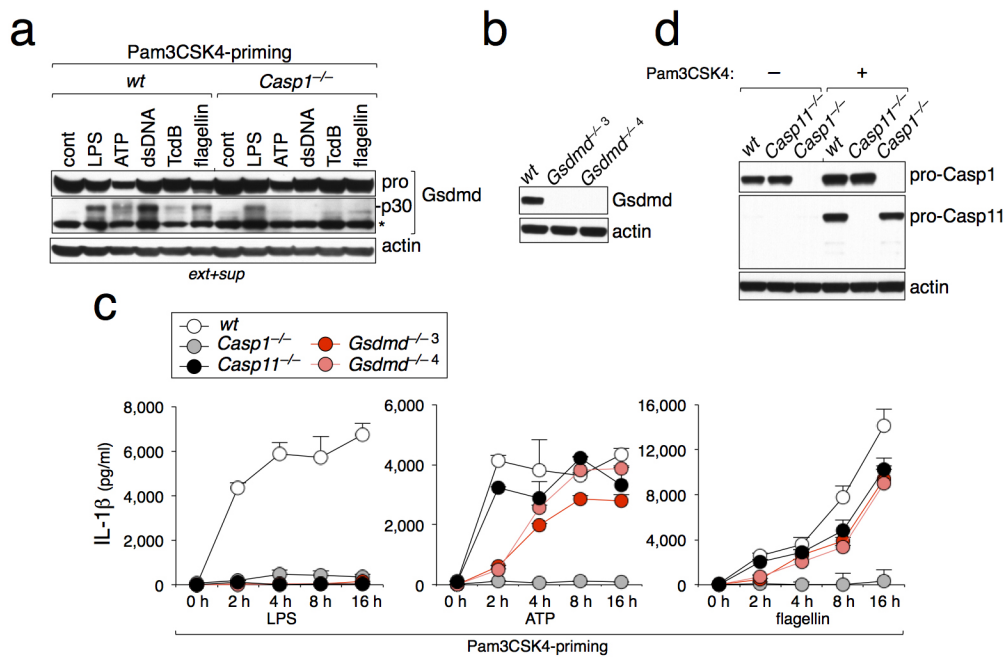
nanograms of plasmid transfected. Western blots (right) indicate protein expression. Expression of the N-terminal fragment (1–275) was below detection levels, presumably due to its potent toxicity. Graph shows mean \pm s.d. of triplicate wells and represents three independent experiments. For source gels of **a**, **b** and **c**, see Supplementary Fig. 3.



Extended Data Figure 6 | BMDMs stimulated with cytoplasmic LPS do not release NLRP3-stimulating damage-associated molecular pattern activity.

a, Paracrine signalling hypothesis. DAMP, damage-associated molecular pattern. **b**, LDH released from BMDMs at 16 h after LPS transfection or electroporation. **c**, IL-1β released from *Casp11*^{-/-} BMDMs at 16 h after

stimulation with ATP or incubation with *Il1b*^{-/-} BMDM culture supernatants derived in **b**. **d**, LDH released from 1:1 mixed cultures of the indicated BMDMs at 16 h after stimulation. Graphs show mean ± s.d. of triplicate wells and represent three independent experiments.



Extended Data Figure 7 | Canonical inflammasome stimuli induce caspase-1-dependent processing of Gsdmd. **a**, Western blots of BMDM extracts and supernatants at 8 h after stimulation. Cont, medium alone. LPS, LPS + Ctb. Asterisk indicates a non-specific band. **b**, Western blots of *Gsdmd*^{-/-} BMDMs. *Gsdmd*^{-/-3} and *Gsdmd*^{-/-4} are independent knockout strains (Extended Data

Fig. 8). **c**, IL-1β released from BMDMs. LPS, LPS electroporation. Graphs show mean ± s.d. of triplicate wells and represent three independent experiments. **d**, Immunoblots of *Casp1*^{-/-} BMDMs stimulated with Pam3CSK4 for 5 h. For source gels of **a**, **b** and **d**, see Supplementary Fig. 3.

Wt gRNA G9 target region

AGTGATGTTGTCAGGCATGGG**GAGAAGGGAAAAATTTCTGGT**GGGGCTGCAGTGTCTGAC
 .ValMetSerLeuGlyMetGlyGluGlyLysIleSerGlyGlyAlaAlaValSerAsp..

***Gsdmd*^{-/-1} (1bpTdel/1bpGins)**

allele1: AGTGATGTTGTCAGGCATGGG**GAGAAGGGAAAAATTTCTGGT**GGGGCTGCAGTGTCTGAC
 allele2: AGTGATGTTGTCAGGCATGGG**GAGAAGGGAAAAATTTCTGGT**GGGGCTGCAGTGTCTGAC

***Gsdmd*^{-/-2} (1bpTdel/1632bpdel)**

allele1: AGTGATGTTGTCAGGCATGGG**GAGAAGGGAAAAATTTCTGGT**GGGGCTGCAGTGTCTGAC
 allele2: -----

***Gsdmd*^{-/-3} (1bpTins/1bpTins)**

allele1: AGTGATGTTGTCAGGCATGGG**GAGAAGGGAAAAATTTCTGGT**GGGGCTGCAGTGTCTGAC
 allele2: AGTGATGTTGTCAGGCATGGG**GAGAAGGGAAAAATTTCTGGT**GGGGCTGCAGTGTCTGAC

***Gsdmd*^{-/-5} (90bpdel+3bpins/10bpdel+5bpins)**

allele1: AGTGATGTTGTCAGGCATGGG**GAGAAGGGAAAAATTTCT**-----
 allele2: AGTGATGTTGTCAGGCATGGG**GAGAAGGG**-----**TGACC**GTGGGGCTGCAGTGTCTGAC

Wt gRNA G2 target region

CGATGGGAACATTC**AGGGCAGAGTGATGTTGTC**AGGCATGGGAGAGGGAAAAATTTTC
 .AspGlyAsnIleGlnGlyArgValMetLeuSerGlyMetGlyGluGlyLysIle...

***Gsdmd*^{-/-4,6-9} (2bpATins/19bpdel)**

allele1: CGATGGGAACATTC**AGGGCAGAGTGATGTTATGTC**AGGCATGGGAGAGGGAAAAATTTTC
 allele2: CGATGGGAACATTC**AGG**-----CATGGGAGAGGGAAAAATTTTC

***Gsdmd*^{-/-10} (2bpATins/2bpATins)**

allele1: CGATGGGAACATTC**AGGGCAGAGTGATGTTATGTC**AGGCATGGGAGAGGGAAAAATTTTC
 allele2: CGATGGGAACATTC**AGGGCAGAGTGATGTTATGTC**AGGCATGGGAGAGGGAAAAATTTTC

***Gsdmd*^{-/-13} (19bpdel/19bpdel)**

allele1: CGATGGGAACATTC**CAGG**-----CATGGGAGAGGGAAAAATTTTC
 allele2: CGATGGGAACATTC**CAGG**-----CATGGGAGAGGGAAAAATTTTC

gRNA G2/G9***Gsdmd*^{-/-11} (G2:1bpTins/G9:5bpdel)**

allele1: CGATGGGAACATTC**AGGGCAGAGTGATGTTGTC**AGGCATGGGAGAGGGAAAAATTTTC
 allele2: AGTGATGTTGTCAGGCATGGG**GAGAAGGGAAAA**-----**GGT**GGGGCTGCAGTGTCTGAC

***Gsdmd*^{-/-12} (G2:203bpdel/G9:4bpdel)**

allele1: -----
 allele2: AGTGATGTTGTCAGGCATGGG**GAGAAGGGAAAAAT**-----**GGT**GGGGCTGCAGTGTCTGAC

Extended Data Figure 8 | *Gsdmd*^{-/-} alleles. *Gsdmd*^{-/-} animals used in this study were compound or homozygous F1 and F2 knockouts generated from mosaic F0 founder and F1 crosses, respectively. gRNA target sequences are highlighted in bold. Deleted bases are indicated by red hyphens. Inserted nucleic acids are highlighted in red.

Extended Data Table 1 | Bioinformatic analysis of ENU-induced SNVs present in IGL1351 pedigree

Chromosome	Coordinate (Assembly version: GRCm38)	Amino Acid Change	Splice Position	Polyphen Score	Polyphen Prediction	SIFT Score	SIFT Prediction	MGI accession Id	Gene Name	Ref Base	Var Base	Amino acid position	Codon
1	75411276		10					MGI:109282	Spag	T	A		
1	80593159	K->N		0.90	Possibly damaging	0.02	deleterious	MGI:2146320	Dock10	T	G	339	TTT
1	151792385	V->A		0.79	Possibly damaging	0.00	deleterious	MGI:1914217	Edem3	T	C	305	GTC
1	190933113	N->K		0.00	Benign	0.28	tolerated	MGI:1196310	Angpt2	C	A	102	AAC
2	119246455	D->G		1.00	Probably damaging	0.01	deleterious	MGI:1338033	Spint1	A	G	340	GAT
3	5401136	Y->C		1.00	Probably damaging	0.02	deleterious	MGI:2137668	Zfhx4	A	G	2118	TAT
3	94886227	Disrupted splicing	6					MGI:1098257	Pamb4	A	T		
3	127672362	Q->R		0.96	Probably damaging	0.00	deleterious	MGI:1918731	Aipk1	T	C	1165	TTG
5	72338951	C->S		1.00	Probably damaging	0.00	deleterious	MGI:1349451	Corin	A	T	673	ACA
6	24734179	F->S		1.00	Probably damaging	0.00	deleterious	MGI:1921659	Hyal6	T	C	37	TTC
7	25295145	R->C		0.98	Probably damaging	0.00	deleterious	MGI:108414	Pafah1b3	G	A	215	ACG
7	35727415	Disrupted splicing	4					MGI:2443952	Dpy19b3	A	G		
7	46761432	M->I		1.00	Probably damaging	0.14	tolerated	MGI:2183207	Hps5	C	T	1113	CAT
7	48867151	T->A		0.12	Benign	0.23	tolerated	MGI:1922038	E2f8	T	C	827	GGT
7	55898243	Y->Stop			N/A			MGI:1338801	Cyfp1	C	A	560	TAC
7	104960739	Disrupted splicing	9					MGI:3030504	Olf670	T	A		
8	25020557	Disrupted splicing	3					MGI:1916992	Tm2d2	A	G		
8	66512896	I->V		0.00	Benign	0.07	tolerated	MGI:1921669	Tkx2	A	G	369	ATC
8	122422123	T->I		0.14	Benign	0.04	deleterious	MGI:2446486, MGI:5141853	Il17c.gm20388	C	T	2	ACC
9	109274572	D->G		0.02	Benign	0.41	tolerated	MGI:1354703	Fbxw14	T	C	347	GTC
10	117291188	C->R		1.00	Probably damaging	0.00	deleterious	MGI:96902	Lyz1	A	G	45	ACA
11	9267565	S->P		1.00	Probably damaging	0.01	deleterious	MGI:2388707	Abca13	T	C	336	TCA
11	70463590	N->K		0.03	Benign	0.90	tolerated	MGI:3603821	Zmynd15	C	A	426	AAC
11	78165400	R->L		0.99	Probably damaging	0.03	deleterious	MGI:1202880	Traf4	C	A	14	CCG
11	110148903	Y->C		0.94	Probably damaging	0.02	deleterious	MGI:2386796	Abca9	T	C	428	GTA
11	119483118	V->A		0.04	Benign	0.18	tolerated	MGI:1289196	Rn213	T	C	5010	GTC
11	121147246	D->E		0.97	Probably damaging	0.11	tolerated	MGI:1920929	Tex19.1	T	A	143	GAT
13	21287499	D->G		1.00	Probably damaging	0.00	deleterious	MGI:104886	Gpx5	T	C	178	GTC
13	38469634	S->P		0.02	Benign	0.21	tolerated	MGI:88182	Bmp6	T	C	226	TCC
14	27464301	E->G		0.06	Benign	0.16	tolerated	MGI:1921694	D14abbb1e	A	G	819	GAG
14	30769820	V->A		0.00	Benign	0.22	tolerated	MGI:1859609	Slmbt1	T	C	17	GTA
14	70145862	T->A		0.99	Probably damaging	0.25	tolerated	MGI:2444228	2610301g19rik	T	C	229	AGT
15	13034240	A->E		0.93	Possibly damaging	0.00	deleterious	MGI:107435	Ctnb6	G	T	778	TGC
15	21237903	V->M		0.99	Probably damaging	0.00	deleterious	MGI:109503	Cdh12	G	A	75	GTG
15	75864337	I->N		0.98	Probably damaging	0.00	deleterious	MGI:1916396	Gsdmd	T	A	105	ATT
17	46726680	D->G		0.01	Benign	0.01	deleterious	MGI:1202304	Gnmt	T	C	124	ATC
18	47281302	K->E		0.23	Benign	0.69	tolerated	MGI:1203727	Sema6a	T	C	520	CTT
18	50281778	W->R		0.00	Benign	0.96	tolerated	MGI:2694939	Fam170a	T	C	164	TGG
19	40607637	I->V		0.00	Benign	1.00	tolerated	MGI:1914840	Tctn3	T	C	309	GAT

Ref Base, reference base. Var Base, variant base.

Extended Data Table 2 | Adjusted *P* values of Fig. 6

Group	-Group	Adjusted p-value
<i>wt</i>	<i>Casp11^{-/-}</i>	0.0002
<i>wt</i>	<i>Gsdmd^{-/-}</i>	0.0002
<i>Casp11^{-/-}</i>	<i>Gsdmd^{-/-}</i>	0.287

Selective small-molecule inhibition of an RNA structural element

John A. Howe^{1*}, Hao Wang^{1*}, Thierry O. Fischmann^{1*}, Carl J. Balibar¹, Li Xiao¹, Andrew M. Galgoci¹, Juliana C. Malinverni¹, Todd Mayhood¹, Artjohn Villafania¹, Ali Nahvi², Nicholas Murgolo¹, Christopher M. Barbieri¹, Paul A. Mann¹, Donna Carr¹, Ellen Xia¹, Paul Zuck³, Dan Riley³, Ronald E. Painter¹, Scott S. Walker¹, Brad Sherborne¹, Reynalda de Jesus¹, Weidong Pan¹, Michael A. Plotkin¹, Jin Wu¹, Diane Rindgen¹, John Cummings¹, Charles G. Garlisi¹, Rumin Zhang¹, Payal R. Sheth¹, Charles J. Gill¹, Haifeng Tang¹ & Terry Roemer¹

Riboswitches are non-coding RNA structures located in messenger RNAs that bind endogenous ligands, such as a specific metabolite or ion, to regulate gene expression. As such, riboswitches serve as a novel, yet largely unexploited, class of emerging drug targets. Demonstrating this potential, however, has proven difficult and is restricted to structurally similar antimetabolites and semi-synthetic analogues of their cognate ligand, thus greatly restricting the chemical space and selectivity sought for such inhibitors. Here we report the discovery and characterization of ribocil, a highly selective chemical modulator of bacterial riboflavin riboswitches, which was identified in a phenotypic screen and acts as a structurally distinct synthetic mimic of the natural ligand, flavin mononucleotide, to repress riboswitch-mediated *ribB* gene expression and inhibit bacterial cell growth. Our findings indicate that non-coding RNA structural elements may be more broadly targeted by synthetic small molecules than previously expected.

Bacterial riboswitches are *cis* regulatory structural elements present in the 5' untranslated region (UTR) of mRNAs that specifically bind to natural ligands and regulate gene expression^{1–6}. Riboswitches are composed of two functionally distinct domains: an aptamer ligand-binding domain, and an expression platform. Distinct families of riboswitches bind specific metabolites, including amino acids, vitamins, glucosamine-6-phosphate, S-adenosylmethionine, thiamine pyrophosphate, ions, and flavin mononucleotide (FMN)^{1–6}. Mechanistically, metabolite binding to the cognate riboswitch aptamer induces a conformational change within the expression platform (hence referred to as a 'riboswitch') leading to altered expression of a gene (or genes) involved in the corresponding biosynthetic pathway^{4,6}. Riboswitch–ligand interactions can therefore provide a negative feedback circuit to modulate metabolism.

Small molecules that are structurally dissimilar to the natural ligand but which mimic its activity to selectively downregulate riboswitch-controlled biosynthetic genes essential for growth (herein referred to as a synthetic mimic) could serve as mechanistically novel therapeutic agents^{7–12}. Riboflavin (vitamin B₂) biosynthesis serves as one potentially attractive metabolic pathway to apply this strategy in antibacterial discovery^{11–14}. FMN riboswitches regulate gene expression of enzymes and transporters of riboflavin biosynthesis and uptake and are broadly conserved across bacterial pathogens yet are absent in vertebrates^{15,16}. Whereas bacteria, fungi, and plants either synthesize this essential vitamin *de novo* (Extended Data Fig. 1) or acquire it from their environment, humans lack the riboflavin biosynthetic repertoire and must consume it as part of their diet. Importantly, riboflavin serves as the immediate precursor to FMN, which is subsequently converted to flavin adenine dinucleotide (FAD), two essential cofactors for a diverse set of flavoenzymes central to intermediary metabolism. Although FMN riboswitches can bind riboflavin and FAD, FMN is the cognate ligand with highest affinity to its receptor to

control riboflavin biosynthesis^{1,11,13}. Therefore, disrupting riboflavin biosynthesis either by synthetic mimics of FMN or enzyme inhibitors of the pathway represents a novel target for antibacterial intervention with limited predicted off-target effects.

To date, little success has been achieved in identifying synthetic mimics of riboswitch ligands despite using a variety of *in vitro* screening strategies, including affinity-based or fragment-based screening, or structure-guided design approaches starting with natural ligands^{7–9}. One of the best-characterized antimetabolite inhibitors to a riboswitch is roseoflavin¹⁷, a naturally-produced analogue of riboflavin, which targets FMN riboswitches from multiple bacterial species^{11,13,18,19}. Like riboflavin, roseoflavin is converted to the flavoenzyme cofactor analogues, roseoflavin mononucleotide (RoFMN) and roseoflavin adenine dinucleotide (RoFAD)¹⁸. In *Bacillus subtilis*, roseoflavin resistant mutants map either to the *ribG* FMN riboswitch or the FAD synthetase gene, *ribC*^{5,20}. Accordingly, the antibacterial effect of RoFMN results from both inhibition of the FMN riboswitch leading to reduced *rib* gene expression¹⁹, as well as inhibiting as many as 40 bacterial flavoenzymes²¹, of which many are conserved in humans^{21,22}. Therefore, roseoflavin or other riboflavin derivatives pose significant potential side-effects as antibacterials and lack the necessary target-selectivity as chemical probes for chemical biology studies.

Here we report a phenotypic screen and the discovery of ribocil, a highly specific bioactive synthetic mimic of FMN, which competes with the natural ligand to inhibit FMN riboswitch-mediated expression of *ribB* and inhibits bacterial growth.

Riboflavin pathway validation and discovery of ribocil

Riboflavin is essential for microbial growth, however, bacterial genes involved in riboflavin biosynthesis are conditionally essential as they are indispensable under conditions where external riboflavin is absent, but not required under conditions where exogenous riboflavin

¹Merck Research Laboratories, Kenilworth, New Jersey 07033, USA. ²Merck Research Laboratories, West Point, Pennsylvania 19486, USA. ³Merck Research Laboratories, North Wales, Pennsylvania 19454, USA.

*These authors contributed equally to this work.

is present and the vitamin can be acquired by either active or passive transport mechanisms^{1,16,23–26}. Accordingly, we first examined the pathogenicity of *Escherichia coli* strains deleted of *ribA* or *ribB*, which correspond to early steps in riboflavin biosynthesis but are viable when grown in the presence of 20 μ M riboflavin supplementation (Extended Data Fig. 1). In a murine septicaemia model, *E. coli* Δ *ribA* and Δ *ribB* mutants are highly attenuated in their virulence, yielding approximately a $3 \times \log_{10}$ reduction in bacterial burden versus the wild-type isogenic control strain at all infectious doses tested (Fig. 1a). Notably, whereas wild-type *E. coli* causes significant mortality at the higher infectious doses tested, mice failed to show signs of morbidity or mortality with Δ *ribA* and Δ *ribB* strains infected at similar or higher infectious inoculates. These results provide genetic evidence that *E. coli* is unable to scavenge sufficient riboflavin in a relevant infection setting and predict that selective inhibitors of riboflavin biosynthesis could display antibacterial efficacy.

We screened an internal library of $\sim 57,000$ synthetic small molecules with antibacterial activity to identify compounds whose growth inhibitory activity against *E. coli* recapitulates the conditional essentiality of the riboflavin biosynthetic pathway. To enhance the opportunity of identifying such molecules, a suitable *E. coli* strain (MB5746) defective in wild-type lipopolysaccharide (LPS) levels and drug efflux was selected²⁷. One compound resulting from this phenotypic screen,

named ribocil (Fig. 1b) demonstrated complete suppression of its bioactivity specifically in the presence of exogenous riboflavin (Fig. 1c). As the conditional growth inhibitory activity of ribocil is a phenocopy of *E. coli* Δ *rib* mutants, it was further investigated as a potential target-specific inhibitor of the riboflavin biosynthetic pathway. Corroborating this view, riboflavin levels in *E. coli* ribocil-treated cells revealed a dose-dependent depletion of riboflavin ($IC_{50} = 0.3 \mu$ M; Fig. 1d), as well as reduced levels of FAD and FMN (Extended Data Fig. 1), mirroring the phenotype of Δ *ribA* and Δ *ribB* mutants (Extended Data Fig. 1).

Ribocil target identification and mechanism of action

To identify the corresponding cellular target of ribocil, we isolated multiple ($n = 19$) independently derived *E. coli* ribocil-resistant (ribocil^R) mutants and mapped drug resistant mutations to their target through whole-genome sequencing (WGS) (Fig. 1e). Remarkably, all drug-resistant mutants contain base pair changes mapping to the FMN riboswitch within *ribB* (Fig. 2a, b). Notably, none of the identified mutations act as compensatory (that is, bypass) mutations conferring drug resistance by dramatically inducing elevated expression of riboflavin either in the presence or absence of ribocil treatment (Extended Data Fig. 1). As WGS failed to detect any additional mutations in all ribocil^R isolates, these data firmly implicate the FMN riboswitch mutations as causal for ribocil resistance (Fig. 1e). The biological significance of this finding is underscored by the essential role of the FMN riboswitch in mediating transcriptional and translational control of the riboflavin pathway²⁵, which controls *ribB* expression via a negative feedback loop in concert with its effector ligand, FMN (Fig. 2a). Based on the functional role of the FMN riboswitch and that all ribocil^R mutations map to this regulatory element, we hypothesized that ribocil inhibits riboflavin biosynthesis directly by mimicking the FMN ligand and binding to the FMN riboswitch to inhibit *ribB* expression.

Direct demonstration that ribocil specifically targets the riboflavin riboswitch was initially achieved using a plasmid-based GFP reporter gene under the control of *E. coli* *ribB* promoter and FMN riboswitch 5' sequence and whose expression is monitored in a ribocil^R isolate, where a ribocil-mediated effect on GFP expression can be quantified without inhibiting cell growth (Extended Data Fig. 2). Ribocil strongly inhibits GFP expression, achieving a 50% effective concentration (EC_{50}) of 0.3 μ M (Extended Data Fig. 2) and paralleling ribocil's 50% inhibitory concentration (IC_{50}) to impair riboflavin synthesis (Fig. 1d). Similarly, ribocil effectively inhibited expression of the GFP reporter when regulated by orthologous riboswitches from *Pseudomonas aeruginosa* or *Acinetobacter baumannii*, albeit at approximately fivefold higher EC_{50} values than against the *E. coli* riboswitch (Extended Data Fig. 2). Conversely, GFP reporter constructs harbouring *E. coli* ribocil^R riboswitch mutations are largely non-responsive to ribocil inhibition (Extended Data Fig. 3), corroborating a direct interaction between ribocil and the riboswitch.

As ribocil lacks potent whole-cell activity against *A. baumannii* and *P. aeruginosa* (see below), classical drug resistance selection studies could not be performed against such organisms. To overcome this issue, we constructed surrogate *E. coli* strains in which the native chromosomal FMN riboswitch element regulating *ribB* was replaced with a functional AbFMN or PaFMN riboswitch. Ribocil^R selections in either heterologous strain background identified 11 and 19 ribocil^R mutations causing unique single nucleotide changes or deletions within the AbFMN and PaFMN riboswitch, respectively (Extended Data Fig. 2). Finally, the frequency of resistance (FOR) to ribocil was quantified against *E. coli* MB5746 and FMN riboswitch recombinant strains (Extended Data Table 1). Whereas the *E. coli* parent (FOR = 2.4×10^{-6}) and PaFMN recombinant (FOR = 6.4×10^{-7}) displayed significant target-based resistance, the AbFMN riboswitch variant yielded an approximately 80-fold lower (FOR = 3.3×10^{-8}) resistance rate compared to the parental strain. Collectively,

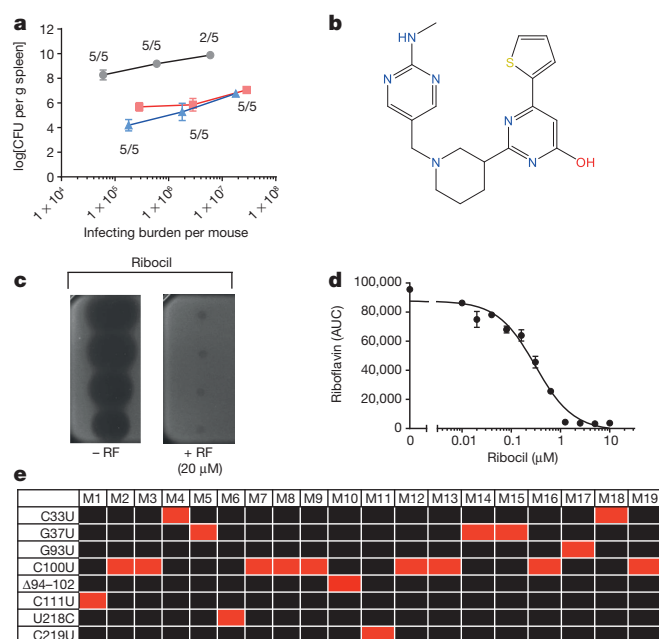


Figure 1 | Genetic and pharmacological inhibition of riboflavin biosynthesis. **a**, Attenuated virulence of Δ *ribA* (red) and Δ *ribB* (blue) mutants versus isogenic parent *E. coli* strain MB5746 (black) across $3 \times \log$ escalation in infectious dose. Colony forming units per gram of spleen were determined at 24 h after infection and are reported as an average ($n = 5$ mice) \pm s.e.m. Numbers denote mouse survival at 24 h time point. **b**, Chemical structure of ribocil. **c**, Whole-cell screening assay to identify riboflavin (RF) inhibitors. In the absence of exogenous riboflavin, ribocil demonstrates a clear zone of growth inhibition when spotted (twofold dilution series) on an agar plate seeded with MB5746. Supplementing the plates with 20 μ M riboflavin fully suppresses ribocil bioactivity. The figure is representative of three independent experiments. **d**, Dose-dependent depletion of cellular riboflavin levels (AUC, area under the curve) by HPLC analysis following ribocil treatment of MB5746 (ribocil $IC_{50} = 0.3 \mu$ M). The data in the figure is the average of two technical repeats (\pm s.d.) and is representative of three independent experiments. **e**, Heat map summary of mutations (in red) identified in MB5746 by illumina-based whole-genome sequencing ($>100 \times$ genome coverage) and verified by Sanger sequencing for 19 independently isolated ribocil^R mutants. All base changes are listed in the left column and map to the FMN riboswitch. Black; no change versus the parental genome sequence.

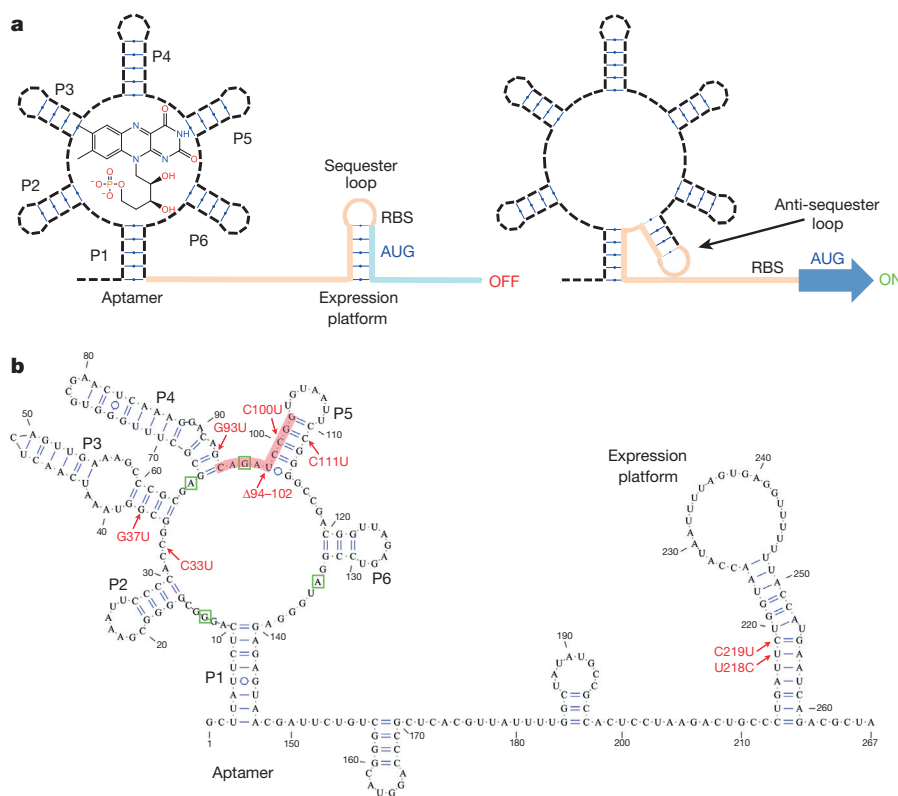


Figure 2 | FMN riboswitch and ribocil^R mutation mapping. **a**, Schematic representation of the FMN riboswitch containing the 5' mRNA FMN aptamer which binds FMN and the 3' expression platform which directly coordinates expression of the downstream *ribB* gene open reading frame (ORF). Left, FMN aptamer binding to the FMN ligand transduces a conformational change in the expression platform, resulting in the formation of a terminator/sequester loop that abolishes *ribB* expression (OFF) in a dual manner: (1) early termination of transcription of the *ribB* ORF; and (2) sequestering the Shine–Dalgarno ribosome binding sequence (RBS) to prevent translation of fully transcribed *ribB* mRNAs²⁵. Alternatively, in the

absence or depletion of FMN (right), the FMN aptamer adopts a second structural conformation (ON) that induces an anti-terminator/anti-sequester loop, thereby facilitating uninterrupted *ribB* expression. In this way, the FMN riboswitch serves as a highly tunable negative feedback system to control riboflavin biosynthesis through the regulation of *ribB* expression. **b**, Predicted secondary structure of the *E. coli* FMN riboswitch and mapping of ribocil^R mutations. Initial secondary structure predictions were derived by global alignment of *E. coli*, *A. baumannii*, *P. aeruginosa* and *F. nucleatum* FMN riboswitches (Extended Data Fig. 5). Ribocil^R mutations are indicated in red. Green boxes indicate predicted key contacts with FMN in the *E. coli* riboswitch.

these genetic data strongly suggest that ribocil directly inhibits FMN riboswitch-mediated *ribB* expression across FMN riboswitches spanning phylogenetically diverse bacteria.

Direct ligand binding of ribocil to the *E. coli* FMN riboswitch was obtained by biophysical means using the synthesized FMN aptamer (Extended Data Fig. 3). To first characterize the structural integrity of the purified aptamer, FMN binding was quantified using spectroscopy to monitor quenching of the intrinsic fluorescence of FMN upon complex formation with RNA^{2,28}. The binding affinity K_d value obtained for FMN binding to the *E. coli* FMN riboswitch aptamer ($K_d = 1.2$ nM), other kinetic parameters, and divalent cation (Mg^{2+}) requirement were found to be similar in range to data reported for FMN riboswitches from a wide variety of bacterial species (Extended Data Fig. 3)^{13,29}. Addition of ribocil in this FMN fluorescence quenching assay restores FMN fluorescence in a potent and dose-dependent manner ($K_d = 16$ nM), demonstrating that ribocil directly competes with FMN for binding to the RNA aptamer (Extended Data Fig. 3).

Structure of ribocil–riboswitch RNA aptamer

Whereas we were unsuccessful in co-crystallizing the *E. coli* FMN aptamer with ribocil, a 2.95 Å structure of ribocil bound to the *Fusobacterium nucleatum* impX FMN riboswitch aptamer was achieved (Fig. 3a, b and Extended Data Fig. 4). The primary sequence of the *F. nucleatum* aptamer conforms well to the *E. coli* sequence as

well as the riboswitch family consensus sequence (Extended Data Fig. 5) and to date is the only FMN riboswitch aptamer for which a crystal structure has been reported¹³. In the ribocil–*F. nucleatum* aptamer co-crystal, ribocil adopts a constrained U-shaped conformation (Extended Data Fig. 4) and, like FMN¹³, is positioned inside the junctional region of the six RNA stems (Fig. 3a). Within the left arm of ribocil, the 6-thiophenyl-pyrimidonyl group stacks face-to-face between A48 and A85 and edge-to-face with A49 (Fig. 3b). A key H-bond also forms between the ribocil pyrimidonyl oxygen and A48 and A99. Overall, the left arm of ribocil binds in a similar conformation as the planar isoalloxazine ring system of FMN (Extended Data Fig. 4), which adopts a continuous stacking alignment bridging the P6 and P3 helices of the riboswitch¹³. Further, ribocil's right arm methyl group at position 5 in the piperidinyl acts as a weak H-bond donor with G11 while the methylamino-pyrimidinyl stacks face-to-face with G62 (Fig. 3b). Notably, both G11 and G62 are conserved guanines that form key interactions with the phosphate group of FMN¹³.

Although a racemic mixture of (S) and (R) enantiomers was used for co-crystallization, electron density mapping of ribocil in the co-crystal structure predicts that only the (S)-isomer is bound (see Methods and Extended Data Fig. 4). Separation of the ribocil mixture led to isolation of the isomers named ribocil-A and ribocil-B. Ribocil-B demonstrated superior microbiological activity as compared to ribocil-A (minimum inhibitory concentration (MIC) = 1 µg ml⁻¹ versus MIC ≥ 64 µg ml⁻¹; Extended Data Table 1), inhibition of

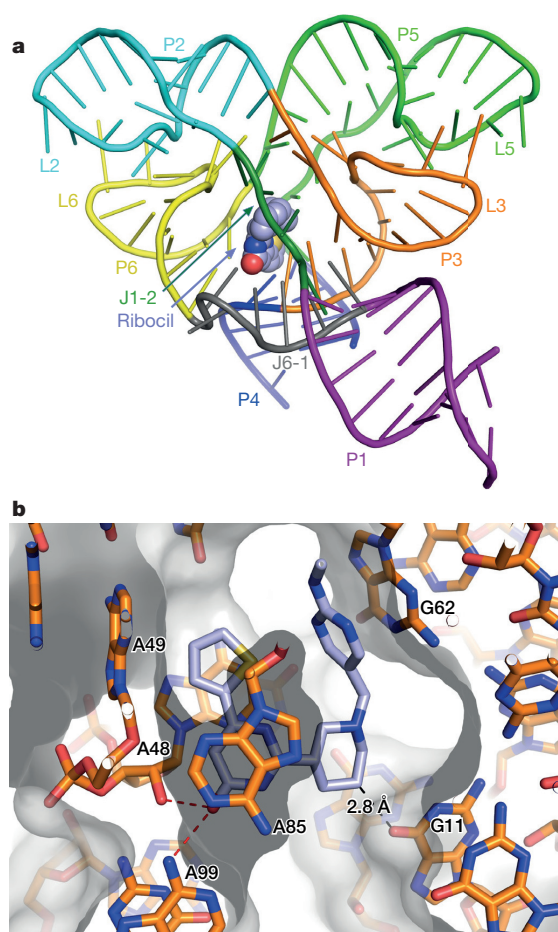


Figure 3 | X-ray crystal structure of ribocil bound to the *F. nucleatum* FMN riboswitch. **a**, Overall structure of the ribocil-bound *F. nucleatum* FMN riboswitch aptamer. The aptamer is represented as a cartoon and coloured by P-loop domain as labelled. The ligand is represented as spheres, with methyls coloured in slate blue, oxygen and nitrogen atoms are in red and blue, respectively. **b**, The RNA and ligand structures are represented as sticks. The carbon atoms coloured in orange for the RNA and slate blue for the ligand. The solvent-accessible surface is shown in grey. The surface is in dark grey where it faces up. Several bases in the vicinity of the ligand are labelled. A85 is positioned on top of the compound in the figure orientation. Polar interactions are shown with red dashes, and an interaction where a methyl acts as weak H-bond donor is drawn as a black dashed line.

riboflavin synthesis ($IC_{50} = 0.13 \mu\text{M}$ versus $IC_{50} > 26 \mu\text{M}$; Extended Data Fig. 3), and binding affinity to the *E. coli* FMN aptamer ($K_d = 6.6 \text{ nM}$ versus $K_d \geq 10,000 \text{ nM}$, Extended Data Fig. 3). Together these results provide compelling evidence that the biologically active ribocil-B isomer is the (S)-isomer bound in the co-crystal.

Analysis of *E. coli* ribocil^R mutations (Fig. 2b, Extended Data Fig. 5 and Supplementary Video) predicts that FMN aptamer mutant $\Delta 94$ –102 directly disrupts ribocil binding by the deletion of a key nucleotide contact (G96) with the compound. This is based on bacterial FMN aptamer homology models which highlight G96 as being equivalent to the G62 nucleotide in the *F. nucleatum* FMN aptamer (Extended Data Fig. 5) and that this guanine stacks with the right arm methylaminopyrimidinyl portion of ribocil in the co-crystal structure (Fig. 3b). Corroborating this view, the EcFMN–GFP fusion containing the $\Delta 94$ –102 allele is unaffected by ribocil ($EC_{50} > 200 \mu\text{M}$) (Extended Data Fig. 3). Alternatively, four other ribocil^R FMN aptamer mutations (C33U, G37U, G93U and C111U) do not directly impair direct contacts with ribocil but probably disrupt Watson–Crick base pairing or tertiary contacts within the FMN riboswitch structure. As such, each of these mutations probably induces a conformational change

within the riboswitch sufficient to effectively reduce ribocil binding affinity. Indeed, all EcFMN–GFP fusions containing these alleles also demonstrate significantly reduced ribocil binding ($EC_{50} \geq 200 \mu\text{M}$) yet preserve sufficient functional expression of *ribB* to maintain cell growth (Extended Data Fig. 3). Alternatively, ribocil effectively inhibits expression ($EC_{50} = 0.2 \mu\text{M}$) of the EcFMN–GFP fusion containing the ribocil^R mutant (C100U), demonstrating that ribocil binding was unaffected and resistance must be conveyed by an alternative mechanism (see below). U218C and C219U map to the expression platform rather than the FMN aptamer (Fig. 2b) and probably uncouple FMN-mediated repression of *ribB* expression by disrupting the requisite base pairing within the sequestration loop to expose the Shine–Delgarno sequence for constitutive *ribB* expression. Consistent with this possibility, ribocil EC_{50} values against EcFMN–GFP containing these mutations are similar to the wild-type EcFMN–GFP fusion ($EC_{50} = 0.2$ – $2.0 \mu\text{M}$) (Extended Data Fig. 3). Although C100U, U218C and C219U bind ribocil with EC_{50} values similar to the wild-type riboswitch, these mutations probably confer ribocil resistance because riboflavin levels are not suppressed sufficiently to inhibit cell growth (Extended Data Fig. 1). Notably, all ribocil^R mutations are markedly cross-resistant to roseoflavin (Extended Data Fig. 6), consistent with ribocil and roseoflavin competitively binding to the FMN aptamer.

Ribocil specifically displays microbiological activity against *E. coli* MB5746 maintaining either the native FMN riboswitch ($MIC = 2 \mu\text{g ml}^{-1}$) or orthologous FMN aptamers (MIC range 1 – $16 \mu\text{g ml}^{-1}$) (Extended Data Table 1) and lacks activity against yeasts and human cells which lack the cognate target (Extended Data Table 1; see Methods). Although ribocil is effectively effluxed in *E. coli* (Extended Data Table 1), inactivity against yeasts is not mediated by efflux (Extended Data Table 1). Ribocil is also more potent than roseoflavin ($MIC > 128 \mu\text{g ml}^{-1}$) against MB5746 (Extended Data Table 1) and, unlike roseoflavin, ribocil bioactivity is fully suppressed by exogenous riboflavin (Extended Data Fig. 6).

To evaluate the effects of inhibiting FMN riboswitch function in a murine *E. coli* septicaemia model of infection, an (S) enantiomer analogue of ribocil named ribocil-C that is mechanistically equivalent to ribocil (Extended Data Fig. 3) and displays eightfold improved antibacterial activity (Extended Data Table 1) was tested. MB5746 bacterial burden in sham-treated mice and low-dose ribocil-C-treated groups (30 mg kg^{-1} ribocil-C) ranged from 9 – $9.5 \log_{10}$ [colony-forming units (CFU) per g spleen] (Fig. 4a,b). Conversely, higher dose ribocil-C treatment groups (60 and 120 mg kg^{-1} ribocil-C) demonstrated a dose-dependent reduction in bacterial burden of 1.87 and $3.29 \log_{10}$ [CFU per g spleen] reduction respectively versus sham-treated mice, without mortality or gross effects of toxicity observed (Fig. 4). These data provide a pharmacological demonstration that riboswitch-mediated inhibition of riboflavin biosynthesis is efficacious in a murine systemic infection model.

Discussion

As a synthetic mimic, ribocil demonstrates high affinity and shares critical contacts to the FMN binding site sufficient to effectively compete with the natural ligand for binding to the riboswitch aptamer and disrupt gene expression. Ribocil elicits its antimicrobial effects in a highly target-selective manner (Supplementary Discussion) and demonstrates that gene regulatory riboswitches are ‘druggable’ by synthetic chemistry. Ribocil also illustrates how such a synthetic mimic demonstrates superior selectivity in its inhibitory effects on whole cells compared to roseoflavin or other antimetabolite ligands. For example, roseoflavin not only targets the FMN aptamer but FAD synthetase²², as well as multiple functionally diverse flavoproteins that require FMN as a cofactor and that are conserved in humans^{21,30}. Conversely, ribocil specifically targets bacterial FMN riboswitches and (unlike roseoflavin) its growth-inhibitory activity is completely suppressed by the addition of exogenous riboflavin (Extended Data Fig. 6); a condition

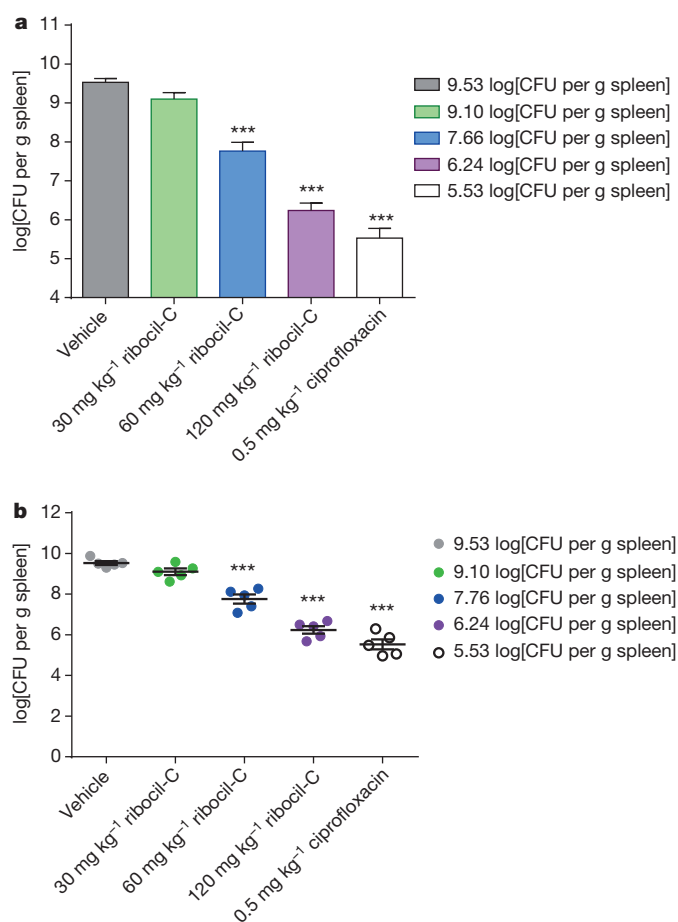


Figure 4 | In vivo activity of ribocil in a murine systemic infection model of *E. coli*. **a**, DBA/2J mice were infected by intraperitoneal injection with *E. coli* strain MB5746 (5.0×10^4 CFU per mouse) and treated by subcutaneous injection with ribocil-C or ciprofloxacin at the indicated doses (mg kg^{-1}) three times over a 24 h infection period. Spleens were aseptically collected from five mice per group and the reduction of log[CFU per g spleen tissue] was calculated on the basis of bacterial burden in spleens of the vehicle-treated (10% DMSO) control group. Data represents the average CFU per g spleen \pm s.e.m. One-way ANOVA with Dunnett's multiple comparison test demonstrates statistically significant ($***P < 0.001$) log[CFU per g] reductions in the 60 and 120 mg kg^{-1} ribocil-C treatment groups and ciprofloxacin control. **b**, The CFU data from each mouse plotted as individual points, solid bar and error bars represent the average CFU per g spleen \pm s.e.m. One-way ANOVA with Dunnett's multiple comparison test demonstrates statistically significant ($***P < 0.001$) log[CFU per g] reductions in the 60 and 120 mg kg^{-1} ribocil-C treatment groups and ciprofloxacin control. Similar efficacy was observed independently at a tenfold higher infectious inoculum (Extended Data Fig. 8).

only expected if ribocil does not additionally inhibit flavoproteins or other essential proteins. Notably, roseoflavin (and presumably other riboflavin antimetabolites) normally requires a highly specific cognate transporter to enter cells that is absent among Gram-negative bacteria^{16,25}. This complicates roseoflavin or similar antimetabolites achieving antibacterial activity against the most serious class of drug-resistant bacterial pathogens³¹. As a synthetic mimic, however, ribocil is not dependent on a riboflavin transport system to enter bacterial cells and inhibit growth.

Phenotypic screens have historically proven highly successful, prompting the renewed interest in them, particularly in the areas of infectious disease and oncology^{32,33}. Notwithstanding potential limitations to the phenotypic screen we describe here (Supplementary Discussion), several advantages to such a screening approach are noteworthy. Importantly, all targets constituting the pathway of interest are screened in an agnostic fashion for cognate inhibitors. Here,

rather than identifying enzyme inhibitors to any of the steps in the riboflavin pathway, ribocil with its unique mechanism of action was identified, exemplifying the value of unbiased phenotypic screens. Importantly, a robust demonstration that target-specific inhibitors are identified in the screen is simply achieved by reversing their bioactivity in a pathway-specific manner; namely, supplementing drug-treated cells with riboflavin. Indeed, suppression of small-molecule bioactivity, whether achieved genetically^{34,35} or through metabolite supplementation³⁶, provides a straightforward means to unambiguously identify bioactive pathway-specific inhibitors. Lastly, performing a chemical similarity search for roseoflavin-like molecules within the screening library revealed several riboflavin analogues not identified in the phenotypic screen. Although all these compounds demonstrate potent antibacterial activity and inhibit riboflavin, FMN and FAD synthesis, none of these compounds are effectively suppressed by riboflavin (Extended Data Fig. 7). Like roseoflavin, we conclude that the growth-inhibiting effects of these riboflavin analogues are also unselective, highlighting the effectiveness of the phenotypic screen to uniquely identify bona fide synthetic mimics of FMN.

Finally, our work suggests 'druggable' targets be considered more broadly from how they are traditionally viewed, which is generally defined as single enzymes, receptors or structural proteins susceptible to chemical inhibition. Indeed, the diverse modes of action that antibacterial agents exhibit emphasizes that therapeutic small molecules interdict substantially more complex targets that are often only relevant in a cellular context³⁷. Although our study emphasizes that target-based drug resistance may compromise the suitability of FMN riboswitches as antibacterial targets, we caution that it is premature to generalize this conclusion across other classes of riboswitches. Indeed, drug resistance may be mitigated by targeting genus-specific riboswitches (Supplementary Discussion) with the added benefit that such antimicrobials may minimally impact gut microbiota³⁸. Finally, human genetic disorders including fragile X syndrome, myotonic dystrophy, and Huntington disease, all result from expanded nucleotide repeats located in transcripts which fold into stable RNA hairpin structures and result in impaired splicing and/or translation³⁹. As such, these disease-causing RNA elements represent a completely novel class of drug targets to which significant drug resistance would not be predicted from somatic cells. It is also estimated that the human transcriptome comprises thousands of non-coding RNAs which also adopt highly specific secondary structures of potential functional and disease-causing significance^{40,41}. Thus, RNA structural elements may substantially expand our view of the target space susceptible to therapeutic intervention.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 29 May; accepted 4 September 2015.

Published online 30 September 2015.

- Winkler, W. C., Cohen-Chalamish, S. & Breaker, R. R. An mRNA structure that controls gene expression by binding FMN. *Proc. Natl Acad. Sci. USA* **99**, 15908–15913 (2002).
- Mironov, A. S. et al. Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria. *Cell* **111**, 747–756 (2002).
- Serganov, A. & Nudler, E. A decade of riboswitches. *Cell* **152**, 17–24 (2013).
- Mellin, J. R. & Cossart, P. Unexpected versatility in bacterial riboswitches. *Trends Genet.* **31**, 150–156 (2015).
- Mack, M., van Loon, A. P. & Hohmann, H. P. Regulation of RF biosynthesis in *Bacillus subtilis* is affected by the activity of the flavokinase/flavin adenine dinucleotide synthetase encoded by ribC. *J. Bacteriol.* **180**, 950–955 (1998).
- Winkler, W. C. & Breaker, R. R. Regulation of bacterial gene expression by riboswitches. *Annu. Rev. Microbiol.* **59**, 487–517 (2005).
- Matzner, D. & Mayer, G. (Dis)similar analogues of riboswitch metabolites as antibacterial lead compounds. *J. Med. Chem.* **58**, 3275–3286 (2015).
- Lünse, C. E., Schüller, A. & Mayer, G. The promise of riboswitches as potential antibacterial drug targets. *Int. J. Med. Microbiol.* **304**, 79–92 (2014).
- Mulhbach, J. et al. Novel riboswitch ligand analogs as selective inhibitors of guanine-related metabolic pathways. *PLoS Pathog.* **6**, e1000865 (2010).

10. Sudarsan, N., Cohen-Chalamish, S., Nakamura, S., Emilsson, G. M. & Breaker, R. R. Thiamine pyrophosphate riboswitches are targets for the antimicrobial compound pyrithiamine. *Chem. Biol.* **12**, 1325–1335 (2005).
11. Lee, E. R., Blount, K. F. & Breaker, R. R. Roseoflavin is a natural antibacterial compound that binds to FMN riboswitches and regulates gene expression. *RNA Biol.* **6**, 187–194 (2009).
12. Blount, K. F. & Breaker, R. R. Riboswitches as antibacterial drug targets. *Nature Biotechnol.* **24**, 1558–1564 (2006).
13. Serganov, A., Huang, L. & Patel, D. J. Coenzyme recognition and gene regulation by a flavin mononucleotide riboswitch. *Nature* **458**, 233–237 (2009).
14. Pedrolli, D. B. *et al.* RF analogs as anti-infectives: occurrence, mode of action, metabolism and resistance. *Curr. Pharm. Des.* **19**, 2552–2560 (2013).
15. Gelfand, M. S., Mironov, A. A., Jomantas, J., Kozlov, Y. I. & Perumov, D. A. A conserved RNA structure element involved in the regulation of bacterial RF synthesis genes. *Trends Genet.* **15**, 439–442 (1999).
16. Vitreschak, A. G., Rodionov, D. A., Mironov, A. A. & Gelfand, M. S. Regulation of RF biosynthesis and transport genes in bacteria by transcriptional and translational attenuation. *Nucleic Acids Res.* **30**, 3141–3151 (2002).
17. Otani, S., Takatsu, M., Nakano, M., Kasai, S. & Miura, R. Roseoflavin, a new antimicrobial pigment from *Streptomyces*. *J. Antibiot. (Tokyo)* **27**, 86–87 (1974).
18. Ott, E., Stolz, J., Lehmann, M. & Mack, M. The RFN riboswitch of *Bacillus subtilis* is a target for the antibiotic roseoflavin produced by *Streptomyces davawensis*. *RNA Biol.* **6**, 276–280 (2009).
19. Pedrolli, D. B. *et al.* A highly specialized flavin mononucleotide riboswitch responds differently to similar ligands and confers roseoflavin resistance to *Streptomyces davawensis*. *Nucleic Acids Res.* **40**, 8662–8673 (2012).
20. Kil, Y. V., Mironov, V. N., Gorishin, I., Krenova, R. A. & Perumov, D. A. Riboflavin operon of *Bacillus subtilis*: unusual symmetric arrangement of the regulatory region. *Mol. Gen. Genet.* **233**, 483–486 (1992).
21. Langer, S., Hashimoto, M., Hobl, B., Mathes, T. & Mack, M. Flavoproteins are potential targets for the antibiotic roseoflavin in *Escherichia coli*. *J. Bacteriol.* **195**, 4037–4045 (2013).
22. Pedrolli, D. B. *et al.* The antibiotics roseoflavin and 8-demethyl-8-amino-RF from *Streptomyces davawensis* are metabolized by human flavokinase and human FAD synthetase. *Biochem. Pharmacol.* **82**, 1853–1859 (2011).
23. Winzler, E. A. *et al.* Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285**, 901–906 (1999).
24. Baba, T. *et al.* Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.* **2**, 2006.0008 (2006).
25. Pedrolli, D. *et al.* The ribB FMN riboswitch from *Escherichia coli* operates at the transcriptional and translational level and regulates RF biosynthesis. *FEBS J.* **282**, 3230–3242 (2015).
26. Eberhardt, S., Richter, G., Gimbel, W., Werner, T. & Bacher, A. Cloning, sequencing, mapping and hyperexpression of the ribC gene coding for riboflavin synthase of *Escherichia coli*. *Eur. J. Biochem.* **242**, 712–719 (1996).
27. Kodali, S. *et al.* Determination of selectivity and efficacy of fatty acid synthesis inhibitors. *J. Biol. Chem.* **280**, 1669–1677 (2005).
28. Wickiser, J. K., Winkler, W. C., Breaker, R. R. & Crothers, D. M. The speed of RNA transcription and metabolite binding kinetics operate an FMN riboswitch. *Mol. Cell* **18**, 49–60 (2005).
29. Rode, A. B., Endoh, T. & Sugimoto, N. Tuning riboswitch-mediated gene regulation by rational control of aptamer ligand binding properties. *Angew. Chem. Int. Edn Engl.* **54**, 905–909 (2015).
30. Langer, S. *et al.* The flavoenzyme azobenzene reductase AzoR from *Escherichia coli* binds roseoflavin mononucleotide (RoFMN) with high affinity and is less active in its RoFMN form. *Biochemistry* **52**, 4288–4295 (2013).
31. Boucher, H. W. *et al.* Bad bugs, no drugs: no ESKAPE! An update from the Infectious Diseases Society of America. *Clin. Infect. Dis.* **48**, 1–12 (2009).
32. Swinney, D. C. & Anthony, J. How were new medicines discovered? *Nature Rev. Drug Discov.* **10**, 507–519 (2011).
33. Moffat, J. G., Rudolph, J. & Bailey, D. Phenotypic screening in cancer drug discovery – past, present and future. *Nature Rev. Drug Discov.* **13**, 588–602 (2014).
34. Swoboda, J. G. *et al.* Discovery of a small molecule that blocks wall teichoic acid biosynthesis in *Staphylococcus aureus*. *ACS Chem. Biol.* **4**, 875–883 (2009).
35. Wang, H. *et al.* Discovery of wall teichoic acid inhibitors as potential anti-MRSA β -lactam combination agents. *Chem. Biol.* **20**, 272–284 (2013).
36. Zlitni, S., Ferruccio, L. F. & Brown, E. D. Metabolic suppression identifies new antibacterial inhibitors under nutrient limitation. *Nature Chem. Biol.* **9**, 796–804 (2013).
37. Walsh, C. Antibiotics: Actions, Origins, Resistance (Washington, DC: ASM Press, 2003).
38. Leffler, D. A. & Lamont, J. T. Clostridium difficile infection. *N. Engl. J. Med.* **372**, 1539–1548 (2015).
39. Guan, L. & Disney, M. D. Recent advances in developing small molecules targeting RNA. *ACS Chem. Biol.* **7**, 73–86 (2012).
40. Carninci, P. *et al.* The transcriptional landscape of the mammalian genome. *Science* **309**, 1559–1563 (2005).
41. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank T. Silhavy for critical reading of the manuscript and providing constructive comments. We also thank the IMCA staff for making the beam line available to us. The X-ray diffraction data were collected by Shamrock (Woodridge, Illinois) and we thank G. Ranieri, J. Carter and R. Walter for collecting the data. We thank L.-K. Zhang (Merck) for helping with HRMS analysis. K. Devito (Merck) and L. E. Smith (Merck) are also thanked for providing cytotoxicity analysis.

Author Contributions T.R. conceived the project; T.R., J.A.H., H.W., T.O.F., C.J.B., A.M.G., J.C.M., T.M., A.N., D.C., J.W., C.G.G., R.Z., P.R.S., C.J.G. and H.T. designed experiments; H.W., J.A.H., T.O.F., C.J.B., L.X., A.M.G., J.C.M., T.M., A.V., N.M., C.M.B., P.A.M., D.C., E.X., P.Z., D.R., R.E.P., S.S.W., B.S., R.d-J., W.P., M.A.P., J.W., D.R., J.C., H.F., performed experiments; T.R. and J.A.H. wrote the manuscript, and all authors analysed data and contributed to editing the manuscript.

Author Information X-ray structure data have been deposited in the Protein Data Bank under accession code 5C45. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to T.R. (terry_roemer@merck.com).

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized.

Microbiological studies. MB5746 is an antibiotic-sensitized *E. coli* strain harbouring an *envA1* mutation and *tolC* deletion⁴². Consequently, MB5746 is outer membrane hyper-permeable and efflux deficient. The primary screen was performed in 1,536-well plate format. MB5746 (1:10,000 dilution of an overnight culture $2-4 \times 10^9$ CFU ml⁻¹) was grown in cation-adjusted Mueller Hinton broth (CAMHB; BD BBL, cat. no. 212322), with or without 10 µM riboflavin (riboflavin; Sigma, cat. no. 4500-100G) supplementation. The final medium volume in each well was 5 µl, to which 50 nl DMSO containing twofold-titrated compound was transferred. Hit compounds whose activity was suppressed by riboflavin were retested with CAMH agar containing MB5746 (1:1,000 dilution of overnight culture). MICs were determined by the broth microdilution method as recommended by the Clinical and Laboratory Standards Institute with one exception: bacterial strains were tested in M9 broth.

Isolation of *E. coli* ribocil resistant mutants. MB5746 was grown to late-exponential phase in CAMHB and spread on CAMH agar plates (BD BBL cat. no. 211438) containing twofold escalating agar MIC levels of ribocil. To establish the number of viable cells in the starting inoculum, the culture was serially diluted and plated on CAMH agar plates lacking ribocil. Resistant isolates were re-streaked on plates containing the same ribocil concentration. The frequency of resistance (FOR) was determined, dividing the number of resistant isolates by the viable CFU in the late-exponential inoculum.

Deletion of *E. coli* *ribA* and *ribB*. *ribA* and *ribB* were knocked out in strain MB5746 by λ-Red recombining using the linear PCR product generated by amplification from pKD4 (ref. 2) with the primer pairs P1/P2 or P3/P4, respectively. Transformation and selection on kanamycin were performed as described previously⁴³, except that 500 µg ml⁻¹ riboflavin was added to the selection plate to maintain the auxotrophic mutants. Upon obtaining strains MB5746Δ*ribA*::*kan* and MB5746 Δ*ribB*::*kan*, riboflavin growth assays were performed to determine the minimal concentration of riboflavin required to maintain growth of the auxotrophic mutant. Liquid cultures of CAMHB inoculated with 1×10^5 CFU ml⁻¹ of the mutants grown overnight on solid media were supplemented with a twofold dilution series of riboflavin ranging from 250 µg ml⁻¹ to 0.25 µg ml⁻¹. After 24 h of incubation at 37 °C growth was scored visually and it was determined that as little as 0.5 µg ml⁻¹ was sufficient to maintain some growth of the auxotrophic mutants, but optimal growth was observed with a minimum of 4 µg ml⁻¹ riboflavin.

RNA preparation of the *E. coli* wild-type FMN aptamers. DNA templates for *in vitro* transcription of aptamer RNA were prepared by PCR from the pCDF-EcFMN-GFP reporter plasmid using a forward primer ApT7 (TAATACGACTC ATTATAGGcttattctcaggcg) incorporating the T7 promoter and a reverse primer ApRev (cggtactctctccatccg). Uppercase represents additional sequences added in the primer including the T7 promoter, lowercase represents the riboswitch sequence. *In vitro* RNA aptamer transcription was carried out using the RiboMAX large scale RNA production system kit (P1300, Promega) using the protocols provided by the manufacturer. After extraction with phenol/chloroform the RNA aptamers were further purified by column chromatography on NAP-10 sephadex columns (GE Healthcare) and isopropanol precipitation.

***In vivo* evaluation of *E. coli* Δ*ribA* and Δ*ribB* virulence.** *E. coli* MB5746 and deletion mutants (Δ*ribA*, Δ*ribB*) were reconstituted in 10 ml trypticase soy broth (TSB, Corning, cat. no. 46-060-CI). The mutant strains were supplemented with 2.5 µg ml⁻¹ riboflavin (Sigma, cat. no. R4500-100G) and incubated at 35 °C for 6 h with shaking at 250 r.p.m. The respective 6-h cultures were used to seed, at a ratio of 1:50 ml, TSB either with or without 2.5 µg ml⁻¹ riboflavin in a 250 ml flask and were incubated at 35 °C for 16 h with shaking at 250 r.p.m. Overnight cultures were centrifuged at 5,000 r.p.m. for 12 min at 5 °C. Supernatant was decanted and pellets were re-suspended in 50 ml fresh TSB to remove excess riboflavin. Six tenfold serial dilutions were made in TSB from these stock cultures (1.2×10^{10} CFU ml⁻¹ for wild type, 5.7×10^9 CFU ml⁻¹ for Δ*ribA*, 3.6×10^9 CFU ml⁻¹ for Δ*ribB*). Select serial dilutions were further diluted 1:10 into 3% gastric hog mucin for intra-peritoneal (i.p.) injection into mice. The initial dilutions without mucin were plated for quantification on TSA II (5% sheep's blood) agar plates (BD BBL, cat. no. 221261) for the wild-type strain or 10 µg ml⁻¹ riboflavin-infused Muller Hinton II Agar (BD BBL, cat. no. 211438) plates for mutant strains.

Eleven-week-old female DBA2/J mice (Jackson Labs) were chosen for this study based on weight (~20 g) and combined into one pool. Animals were then randomly selected from this pool and placed in groups of five in separate boxes. Subjects were treated with 150 mg kg⁻¹ i.p. cyclophosphamide (Baxter, NDC# 10019-955-50) on day -4 and 100 mg kg⁻¹ on day -1. On day 0, five mice per

group were injected i.p. with 0.5 ml of a respective dilution of bacteria in 3% mucin (6.0×10^6 , 10^5 , 10^4 CFU ml⁻¹ for wild type, 2.85×10^7 , 10^6 , 10^5 CFU ml⁻¹ for Δ*ribA*, 1.8×10^7 , 10^6 , 10^5 CFU ml⁻¹ for Δ*ribB*). On day 1, subjects were euthanized via CO₂ asphyxiation and spleens were aseptically removed, weighed and homogenized in 1.5 ml of sterile saline (Hyclone, cat. no. SH30028.03) with 10% glycerol (Fisher Scientific, cat. no. BP229-1). Tissue homogenates were serially diluted tenfold in sterile saline and selected concentrations were plated on either TSA II (5% sheep's blood) agar plates for the wild type or MH riboflavin infused agar plates for Δ*ribA* and Δ*ribB* mutants. Plates were incubated at 35 °C for 24 h and CFU per g of spleen tissue were determined. No data was excluded from this study and investigator blinding was not implemented during this study. This study was approved and was in compliance with the ethical regulations set forth by the Institutional Animal Care and Use Committee (IACUC) at Merck Research Laboratories, Kenilworth, New Jersey.

Antibacterial effect of ribocil-C in a mouse septicemia model. *E. coli* MB5746 was reconstituted in 10 ml trypticase soy broth (TSB, Corning, cat. no. 46-060-CI) and incubated at 35 °C for 6 h with shaking at 250 r.p.m. The 6 h culture was used to seed, at a ratio of 1:50 ml, TSB in a 250 ml flask and were incubated at 35 °C for 16 h with shaking at 250 r.p.m. The overnight culture was centrifuged at 5,000 r.p.m. for 12 min at 5 °C. Supernatant was decanted and the pellet was re-suspended in 50 ml fresh TSB to remove excess riboflavin. Nine tenfold serial dilutions were made in TSB from the culture (1.0×10^{10} CFU ml⁻¹). The third dilution (1.0×10^7 CFU ml⁻¹) was further diluted into 3% gastric hog mucin for i.p. injection into mice. The initial dilutions without mucin were plated for quantification on TSA II (5% sheep's blood) agar plates (BD BBL, cat. no. 221261).

Twelve-week-old female DBA2/J mice (Charles River Laboratory) were chosen for this study based on weight (~20 g) and combined into one pool. Animals were then randomly selected from this pool and placed in groups of five in separate boxes. Subjects were treated by intraperitoneal (i.p.) injection with 150 mg kg⁻¹ of cyclophosphamide (Baxter, NDC# 10019-955-50) on day -4 and 100 mg kg⁻¹ on day -1. On day 0, mice were inoculated i.p. with 0.5 ml of bacteria in 3% mucin (5.0×10^4 CFU ml⁻¹; Fig. 4) or a higher inoculum of 5.0×10^5 CFU ml⁻¹ (Extended Data Figure 8). Thirty minutes post-inoculation, mice ($n = 5$ per group) were treated by subcutaneous (s.c.) injection three times over 24 h with either ciprofloxacin (0.5 mg kg⁻¹, Sigma Aldrich, cat. no. 17850-5G-F, ribocil-C (at either 120, 60, or 30 mg kg⁻¹)) or 10% DMSO (Sigma Aldrich, cat. no. 276855-1L) sham. On day 1, subjects were euthanized via CO₂ asphyxiation and spleens were aseptically removed, weighed and homogenized in 1.5 ml of sterile saline (Hyclone, cat. no. SH30028.03) with 10% glycerol (Fisher Scientific, cat. no. BP229-1). Tissue homogenates were serially diluted tenfold in sterile saline and selected concentrations were plated on TSA II (5% sheep's blood) agar plates. Plates were incubated at 35 °C for 24 h and CFU per g of spleen tissue were determined. A normality test was performed to verify normal distribution of data before determining statistical significance via the one-way Bonferroni ANOVA. No data was excluded from these studies and investigator blinding was not implemented during this study. This study was approved and was in compliance with the ethical regulations set forth by the Institutional Animal Care and Use Committee (IACUC) at Merck Research Laboratories, Kenilworth, New Jersey.

HPLC-based quantitative analysis of flavins. Overnight cultures of MB5746 or MB5746 Ribocil^R cells were diluted 1:50 in CAMHB and distributed (1.25 ml) into 10-ml culture tubes containing diluted ribocil (twofold dilution series) or DMSO (1%) as mock control. The treated cultures were incubated with shaking at 37 °C for about 20 h, after which the OD₆₀₀ of the culture was determined and 500 µl was moved to a 96-well deep-well plate. After centrifugation (4,000 r.p.m.) for 10 min, the bacterial cell pellets were rinsed with lysozyme dilution buffer (10 mM Tris HCl (pH 8.0), 25 mM NaCl, 1 mM EDTA) and centrifuged again. Cell pellets were then re-suspended in 100 µl of lysozyme solution (10 mg ml⁻¹ lysozyme (Sigma) in lysozyme dilution buffer), incubated at 37 °C for 30 min, and then frozen at -20 °C. Riboflavin, FMN and FAD concentrations in the bacterial lysates were determined using the Vitamin B2 HPLC detection kit (ImmuChrom, GmbH) and Vitamin B2 column (IC2300rp, ImmuChrom GmbH) following the procedure recommended by the manufacturer scaled for a 50 µl sample (bacterial lysate). A Shimadzu HPLC system with fluorescence detector was used at a flow rate of 1.0 ml min⁻¹ and flavin detection was carried out at 450 nm. Flavin levels were determined for an equivalent number of cells by correcting raw AUCs using the OD₆₀₀ ratio of the treated versus the untreated cultures.

Oligonucleotides used for reporter and strain construction. P1:tatggcaaaat aagcaatacagaaccagcattatctggagaattcatgtgctgaggctggagctcttc; P2:aagcaaatgaattacacatgcaagggttattttgtcagcaatggcccatgaatatcctcttag; P3:gactgcccgtattctggta accataatttagtgagggtttttaccatgtgtgaggctggagctcttc; P4:gattaaaggcagtaaatgaagcagc ggttttcagctggctttacgctcatgatgaatatcctcttag; P5:CTCAATGCCTGAGGTTT CAGcaggactgtcgtttggacgtc; P6:GAAAAGTCTCTCTCTTACTCATgtgtaaaaa

acctcactaaaattatg; P7:GACGTCCAAACGCAAGTCTCTGctgaacctcaggcttag; P8:CATGGATGAGCTCTACAAATAAGcgaacgaataatgtaag; P9:CATAATTT TAGTGAGGTTTTTTTACCatgagtaagaggaagaaacttttc; P10:CTTACATTAATT GCGTTGCGCttattgttagactcatcatg; P11:CATTAGCGTTATAGTGAATCCGC taactgtctcaggcgggggtg; P12:GAAAAGTTCTTCTCTTTACTCATgcaactccgtttt tccgcc; P13:CATTAGCGTTATAGTGAATCCGCtaaaacctcgttcagggg; P14:GA AAAGTTCTTCTCTTTACTCATaatgaacctctcgtgaagaatac; P15:CACCCCGC CTGAGAACGTTAgcggtactcataacgctaag; P16:GGCGGAAAAACGGGAGGT CGCatgagtaaggaagaagaacttttc; P17:GCCCTGAAGCGATGGGTTTTAgcggtac tataacgctaag; P18:GTATTCTTACGAGAGCGTTTCATttagtaaaaggagaagaact ttc; AbRFN-ribB_RED forward: ttgcatcgtctgaaatgttcaacgctattcttagagagcggttcatta tgaatcacagcgtactttc; PaRFN-ribB_RED forward: gtcgcgcggcgatcgtgcgcgtctgtcg gcgcaaaaacgggaggtcgcatgaatcagcgtactttc; Ab/Pa RFN-ribB_RED reverse: agatccc ggtgcctaagtgagctaaactcataatgctgtgcgcgtggtttacgtctatg; yqiC_RED forward: caggacttgcgttggagcgtcgaactctcagcgcttaacagtcaggcgctcagctgcgtttagg; ribB_RED reverse: gattaaggcagtaaatgaagcagcggtttcagctggtttacgtctatgctgcgtgag gtagtcacca; yqiC seq. reverse: agtgcgtgattctgttc; ribB seq. reverse 1: agcgaatt aacatcttg; ribB seq. reverse 2: gcttcaatggtcaggtaa. Transition from capital to lowercase letters for P5–P18 denotes boundaries of fragments that facilitate in-fusion cloning.

FMN–GFP reporter construction. EcFMN–GFP reporter plasmids were constructed by fusing the EcFMN region, inclusive of 550 bp upstream of *ribB* through the start codon, to *gfpuv* and cloning into a vector with the low copy CloDF13 origin of replication. Primers P5 and P6 were used to amplify the EcFMN region from wild-type and resistant mutants by colony PCR, primers P7 and P8 were used to amplify the CloDF13 origin and *Sm^R* cassette from pCDF-1b (EMD Millipore), and primers P9 and P10 were used to amplify *gfpuv* from pGFPuv (Clontech). Upon purification, all three linear PCR products were combined using the in-fusion HD cloning system (Clontech) and transformed into TOP10 cells (Life Technologies) with selection on spectinomycin (MP bio-medicals) to yield pCDF-EcFMN–GFP reporter plasmids. Plasmids were subsequently transformed into the MB5746 ribocil^R mutant M5 background for compound testing.

Initial attempts to create *Pseudomonas aeruginosa* and *Acinetobacter baumannii* FMN–GFP reporters in a similar fashion to the EcFMN–GFP reporters by using the PaFMN or AbFMN region, inclusive of 550 bp upstream of *ParibE* or *AbribB* through the start codon, did not yield constructs with sufficient baseline fluorescence (data not shown). In order to optimize fluorescence, hybrid constructs were made in which the *E. coli* promoter region was placed upstream of the PaFMN or AbFMN elements. Primer combinations P11/P12 or P13/P14 were used to amplify the PaFMN or AbFMN elements, respectively, from wild-type cells by colony PCR, and primer pairs P15/P16 and P17/P18 were used to amplify the *E. coli* promoter, *gfpuv*, and vector backbone from the previously constructed pCDF-EcFMN–GFP plasmid for combination with PaFMN and AbFMN, respectively. Purified linear PCR products were combined and transformed as described above to yield pCDF-EcPro-PaFMN–GFP and pCDF-EcPro-AbFMN–GFP reporter plasmids. Again, plasmids were transformed into the MB5746 ribocil^R mutant M5 background for compound testing.

Construction of the hybrid AbFMN and PaFMN *EcribB* strains. The native, chromosomal *E. coli* *ribB* riboswitch was replaced with that of either *A. baumannii* or *P. aeruginosa* using a two-step λ -RED recombineering process⁴⁴. In the first recombineering event, the GFPuv coding sequence from either the pCDF-EcPro-AbFMN–GFP or pCDF-EcPro-PaFMN–GFP plasmid was replaced with the *E. coli* *ribB* coding sequence (*EcribB*). To this end, MB5746 *ribB::kan* cells were grown in CAMH broth supplemented with 4 $\mu\text{g ml}^{-1}$ riboflavin (reconstituted in 1:1 dH₂O:ethanol) and transformed (as described below) with the temperature-sensitive plasmid pKD46⁴⁴. Reactions were plated onto CAMH agar + 50 $\mu\text{g ml}^{-1}$ ampicillin at 30 °C. Next, either the pCDF-EcPro-AbFMN–GFP or pCDF-EcPro-PaFMN–GFP plasmid was transformed into MB5746 *ribB::kan/pKD46* and plated onto CAMH + 4 $\mu\text{g ml}^{-1}$ riboflavin + 50 $\mu\text{g ml}^{-1}$ spectinomycin + 50 $\mu\text{g ml}^{-1}$ ampicillin at 30 °C to maintain double plasmid selection. The resulting strains were recombineered with *EcribB* PCR product containing flanking regions homologous to the cognate pCDF plasmid. Substrate PCR products were obtained through colony PCR of the wild-type *ribB* locus of MB5746 and either the AbFMN-ribB_RED forward or PaFMN-ribB_RED forward primer in combination with the Ab/Pa FMN-ribB_RED reverse primer. Cells were recovered in CAMH broth for 1 h and Rib⁺ colonies were selected on CAMH agar + 50 $\mu\text{g ml}^{-1}$ spectinomycin and incubated at 37 °C to remove pKD46. The resulting strains, MB5746 *ribB::kan* (pCDF-EcPro-AbFMN-*EcribB*) or (pCDF-EcPro-PaFMN-*EcribB*) carry a plasmid-borne *EcribB* gene downstream of the native *E. coli* *ribB* promoter fused to either AbFMN or PaFMN.

In the second recombineering event, the plasmid-borne AbFMN- or PaFMN-*EcribB* fusions engineered above were introduced into the *E. coli* chromosome in

single-copy at the native *ribB* locus. MB5746 *ribB::kan/pKD46* cells were grown to exponential phase in CAMH broth + 4 $\mu\text{g ml}^{-1}$ riboflavin + 50 $\mu\text{g ml}^{-1}$ ampicillin and electroporated with either AbFMN-*EcribB* or PaFMN-*EcribB* PCR product containing flanking regions homologous to the native *ribB* locus. These PCR products were amplified from pCDF-EcPro-AbFMN-*EcribB* or pCDF-EcPro-PaFMN-*EcribB* using the yqiC_RED forward and ribB_RED reverse primers. Unlike strains carrying the plasmid-borne hybrid *EcribB* constructs, the chromosomal hybrid fusion constructs do not yield enough riboflavin for optimal growth on CAMH in single copy. Therefore, reactions were recovered in CAMH broth containing very low levels of riboflavin (0.4 $\mu\text{g ml}^{-1}$, a concentration that does not permit growth of the *ribB* deletion mutant), plated onto CAMH agar + 0.4 $\mu\text{g ml}^{-1}$ riboflavin, and incubated at 37 °C to remove the pKD46 plasmid. The *E. coli* *ribB* promoter, hybrid riboswitch, and *EcribB* coding regions were sequenced in resulting Rib⁺ cells and additionally sequenced at joint sequences using yqiC seq. reverse, ribB seq. reverse 1, and ribB seq. reverse 2 primers.

All transformations were electroporation reactions performed as suggested² with some modifications. Around 30–50 ml of cells were grown in CAMH to exponential phase. For recombineering reactions, strains harbouring pKD46 were induced for 1 h with 1% arabinose before harvesting of cells. Cells were washed with 30 ml of ice-cold ddH₂O and pelleted at 4 °C, 3,000g for 10 min, followed by two additional washes with 1 ml ice-cold ddH₂O and pelleted each time at 8,000g at 4 °C for 2 min. Pellets were re-suspended in 300 μl ddH₂O and 100 μl cells were incubated with 1–2 μl PCR product for 5 min before electroporation. Electroporation reactions were performed using 0.1-cm gap cuvettes and a GenePulser II (BioRad) with settings at 200 Ω , 25 μF , and 1.8 kV. Cells were recovered at the appropriate temperature in 1 ml CAMH broth as described above and plated on CAMH agar containing the appropriate supplements. All PCR reactions were performed using GoTaq Green Master Mix (Promega Corporation) according to manufacturer's instructions. DNA sequencing was performed by Genewiz, Inc.

Crystallization and RNA–ribocil X-ray co-structure determination. Crystals of the *F. nucleatum* FMN riboswitch in presence of the ligand were obtained following published protocols^{13,45} with minor modifications. The RNA was synthesized in two strands: GGAUCUUCGGGGCAGGGUGAAAUCCCCGACCGG UGGUAUAGUCCACGAAAGCUU and GCUUUGAUUUGGUGAAAUCC AAAACCGACAGUAGAGUCUGGAUGAGAGAAGAUUC. The oligonucleotides were purchased from Sigma-Aldrich. After reception each strand was dissolved in water, aliquoted so that each aliquot would contain the material necessary to make a 25 μl solution at 0.4 mM concentration. The aliquots were lyophilized using a Centriva concentrator (Labconco) and kept at –20 °C for long-term storage. Prior to annealing the nucleic acids were re-suspended in 25 μl annealing buffer (10 mM cacodylic acid, 100 mM acetate, 4 mM MgCl₂ adjusted to pH 6.8 using KOH). The oligomers were mixed together, along with 1.0 μl of an inhibitor stock solution at 50 mM in 100% deuterated DMSO, and annealed in a thermocycler by incubation at 37 °C for 30 min followed by cooling from 37 °C to 4 °C at a rate of 3 °C per min. The crystals were grown by vapour diffusion using a 15-well EasyXtal DropGuard X-Seal tool (Qiagen) after mixing 3 μl of riboswitch–ligand solution with 3 μl precipitant (0.1 M Na acetate, pH 5.0, 0.2 M MgCl₂, and 7 to 11% v/v PEG 4K). Small nuclei appear after a few days, and are made to grow larger for diffraction studies by controlled drying. Drying is achieved by substitution once a day of the adequate volume of well solution with a 50% v/v PEG 4K stock solution. The volume is calculated to achieve a ~2% increase in precipitant concentration per step. The crystals after growth are harvested and dipped for 1 to 2 min in a cryoprotectant solution (0.1 M Na MES, pH 6.5, 0.2 M MgCl₂, 10% v/v PEG 4K, 20% v/v glycerol, and ligand diluted to 1 mM concentration). Crystals were harvested with a mesh Litholoop (Molecular Dimensions Ltd) and flash-frozen in liquid nitrogen.

X-ray crystal structure determination. X-ray diffraction data (Extended Data Table 2) were collected at the Advanced Photon Source (APS) sector 17 (IMCA) at 1.0 Å wavelength using a Pilatus 6M (Dectris) pixel array detector. 720 frames with an oscillation of 0.25° each were collected. The data were processed using the automated pipeline autoPROC⁴⁶, with calls to the programs XDS⁴⁷ for integration and AIMLESS⁴⁸ for scaling. The structure was determined using PDB entry 3F4E as a starting point after removing all heterogeneous atoms including the FMN. The structure was refined without inclusion of the ligand coordinates at any step before and until an omit map difference map is generated to fit the compound. The steps include refinement using the program autoBUSTER⁴⁹, corrections of the model and inclusions of several cations with Coot, and Cartesian simulated annealing using the program Phenix⁵⁰ to further eliminate the potential of bias against FMN which was present in the PDB 3F4E entry. The set of 'free' reflections was taken from the same PDB entry 3F4E and completed as required. All refinement calculations after adding the ligand were performed using the

program autoBUSTER⁴⁹. Model visualization and rebuilding was performed using the program Coot⁵¹. All figures in the manuscript generated with PyMol⁵². **Compound chirality.** Co-crystallization of the heptamer with ribocil was performed using a racemic mixture of the ligand. In spite of its limited resolution, 2.9 Å, the (*R*) isomer fits distinctly better than the (*S*) isomer in the initial electron density. Consistent with this observation, crystallographic refinement of a model which starts with the wrong (*R*) isomer ends up with a structure with the chiral centre inversed and nearly planar, an impossible stereochemistry. By contrast, the chiral volume remains unchanged in the course of refinement when starting from the (*S*) isomer. Notwithstanding the electron density map, due to the constrained nature of the binding site it is not possible to fit the (*R*) isomer while maintaining reasonable ligand stereochemistry and parallelism of the pyrimidinonyl and the methylaminopyrimidinyl between the bases of A48 and A85, and against the base of G62. Altogether, crystallographic and stereochemical considerations strongly support the conclusion that only the (*S*) form binds to the FMN aptamer. Further observations made later when the isomers were separated agree with this interpretation: only one of them is active against the riboswitch, and the ligand with the correct chirality is more active than the racemic mixture (Extended Data Fig. 3 and Extended Data Table 1).

Homology modelling. A homology model of the *E. coli* FMN aptamer was constructed using program mutate_bases⁵³ of the 3DNA package using the *F. nucleatum* impX riboswitch aptamer X-ray structure as the template and the FMN aptamer alignment of *E. coli*, *F. nucleatum*, *P. aeruginosa* and *A. baumannii* (Extended Data Fig. 5). All nucleotide insertions in the *E. coli* sequence were removed in the model (Extended Data Fig. 5). There are 34 base changes among the 111 nucleotides modelled. Base pairing when present remains consistent. Energy minimization at A92 was performed to avoid VDW clashes using MacroModel (Schrodinger, LLC).

FMN–GFP reporter assays. Reporter strains were diluted to $\sim 5 \times 10^6$ CFU ml^{−1} in CAMHB supplemented with 30 µg ml^{−1} spectinomycin. Compounds to be tested were serially diluted twofold through 11 points in 100% DMSO. A BioMek FX liquid handler was used to deliver 49 µl of diluted culture into a 384-well, black/clear-bottom assay plate followed by 1 µl of titrated compound. DMSO and antibiotic controls were added manually to appropriate wells and the plates were shaken for 1 min before incubating at 37 °C. After overnight growth, fluorescence, using 405 nm excitation and 510 nm emission, and absorbance at 600 nm, was assessed on an EnVision multiplate reader (Perkin Elmer). Fluorescence response (RFU), relative to full growth and fully inhibited (50 µM ribocil) controls and absorbance response (OD₆₀₀), relative to full growth and sterile controls, were fitted to four parameter (variable slope) curves. The concentration of compound which decreased the specific fluorescence signal by 50% is reported as the GFP EC₅₀.

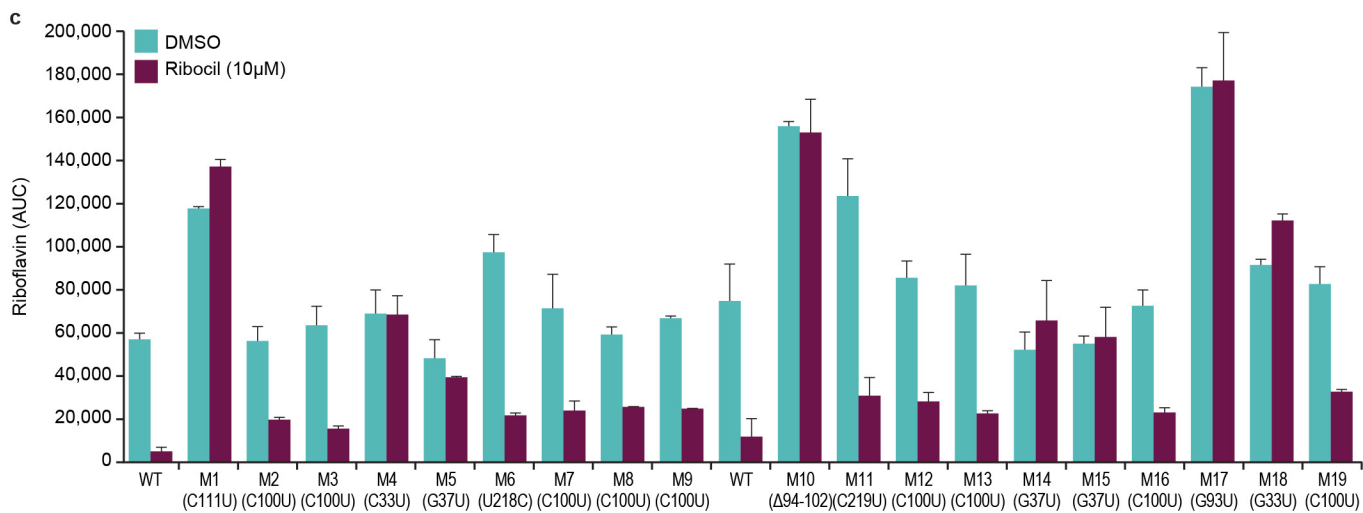
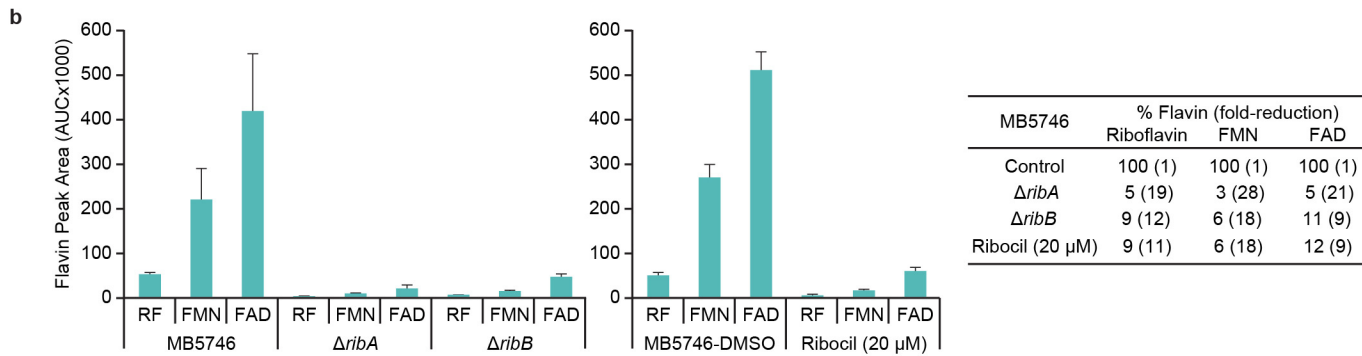
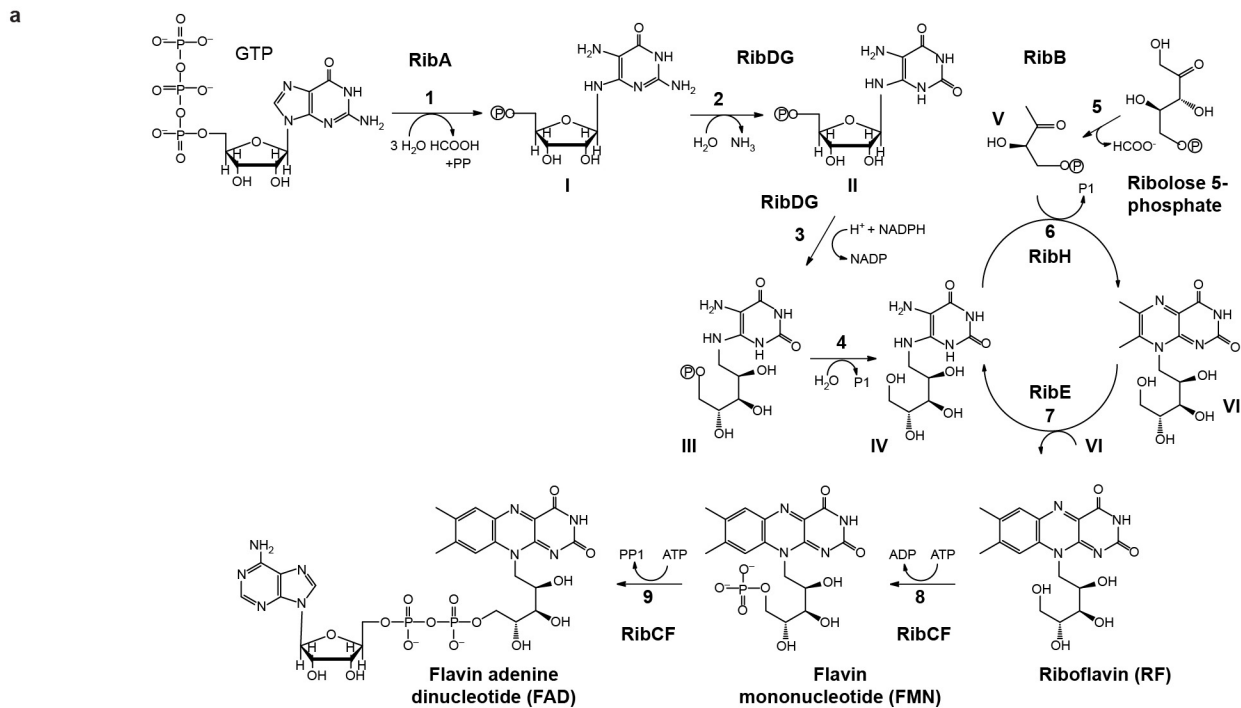
FMN binding to riboswitch aptamer. Aptamers were first re-annealed at a 20 µM concentration in 4 mM KH₂PO₄, 16 mM K₂HPO₄, 64 mM KCl and 0.1 mM EDTA, pH 7.4 buffer by heating at 95 °C for 5 min followed by incubation at room temperature for 15 min. Only one re-annealing cycle was performed per aptamer sample. A 1.25-fold serial dilution of the re-annealed aptamer was prepared to have a final concentration ranging from 6.6 to 150 nM in 50 mM Tris-HCl, 100 mM KCl and 2 mM MgCl₂ assay buffer, pH 7.4. This was mixed with FMN ligand (30–240 nM final concentration). Fluorescence signal was read using the Spectramax M5 at an excitation wavelength of 455 nm and emission wavelength at 525 nm with cut-off filter at 515 nm. The instrument was set up in kinetic mode to acquire data every 20 s. The steady-state *K*_d and the binding-competent fraction of aptamer were determined from fluorescence data obtained at 120 min

of the reaction by fitting to a quadratic equation fully describing the binding equilibrium under tight-binding conditions.

Competition binding kinetics. A twofold serial dilution of compounds was prepared to have a final assay concentration range from 1.22 to 10,000 nM. This was prepared in 50 mM Tris-HCl, 100 mM KCl and 2 mM MgCl₂ assay buffer, pH 7.4, with 0.2% DMSO. FMN ligand concentration was 60 nM and the *E. coli* FMN aptamer concentration was 48 nM or, for ribocil, 150 nM. The fluorescence signal was read on the Spectramax M5 as described above. The steady-state binding competition data at 120 min was fitted to a cubic equation fully describing the competition binding equilibria to derive the *K*_i value for the compound, while fixing *K*_d to the value obtained earlier for FMN binding. The binding kinetics data was fitted by KinTek Explorer-based numerical integration with *K*_i constrained to derive the dissociation rate constant (*k*_{off}). The association (*k*_{on}) rate constant is then calculated from *K*_i and *k*_{off}.

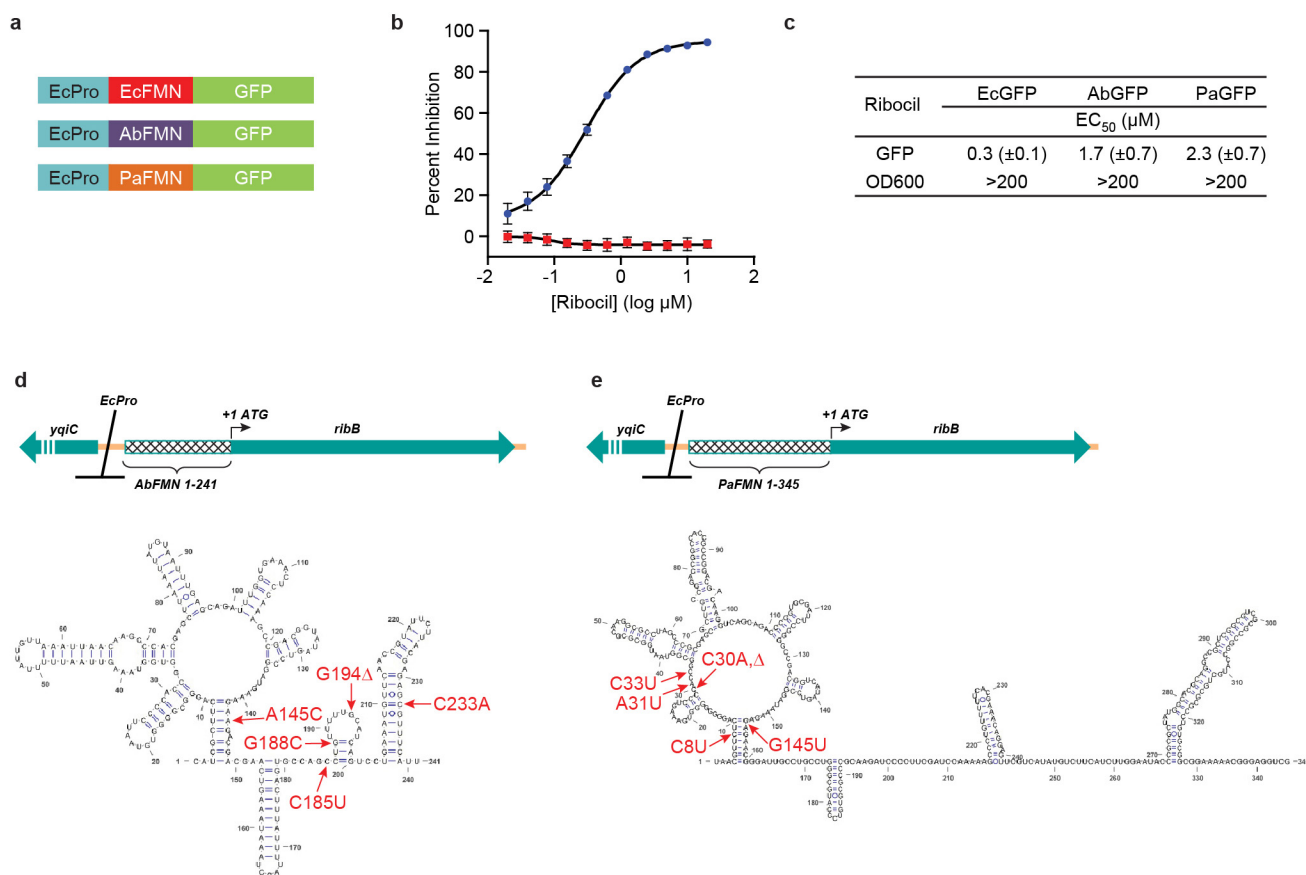
Mammalian cytotoxicity assay. The Click-iT EdU Alexa Fluor 488 HCS assay kit (Life Technologies, C10351) was used to assess the potential cytotoxicity of ribocil in mammalian cell cultures using a modified version of the manufacturer's protocol. For the assay, mycoplasma-tested HeLa cells (ATTC) were seeded at 4,000 cells per well in 384-well poly-D-lysine-coated plates (Greiner, 781946) in 25 µl of culture medium (Optimem I, Life Technologies) and treated with a 20-point twofold dilution series of ribocil. After addition of EdU (5 µM) and incubation (37 °C) for 24 h, images were captured and analysed using an Acumen eXC3 (TTP Labtech Ltd) laser scanning cytometer. Total cell numbers were determined using Hoechst 33342 (Life Tech, H3530). Ribocil displayed no HeLa cell cytotoxicity as detected by cell count (EC₅₀ ≥ 100 µM, activity at 100 µM = 17%) or EdU measurement (EC₅₀ ≥ 100 µM, activity at 100 µM = 11%).

42. Kodali, S. *et al.* Determination of selectivity and efficacy of fatty acid synthesis inhibitors. *J. Biol. Chem.* **280**, 1669–1677 (2005).
43. Balibar, C. J., Hollis-Symynkywicz, M. F. & Tao, J. Pantethine rescues phosphopantothenoylcysteine synthetase and phosphopantothenoylcysteine decarboxylase deficiency in *Escherichia coli* but not in *Pseudomonas aeruginosa*. *J. Bacteriol.* **193**, 3304–3312 (2011).
44. Datsenko, K. A. & Wanner, B. L. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc. Natl Acad. Sci. USA* **97**, 6640–6645 (2000).
45. Vicens, Q., Mondragón, E. & Batey, R. T. Molecular sensing by the aptamer domain of the FMN riboswitch: a general model for ligand binding by conformational selection. *Nucleic Acids Res.* **39**, 8586–8598 (2011).
46. Vonrhein, C. *et al.* Data processing and analysis with the autoPROC toolbox. *Acta Crystallogr. D* **67**, 293–302 (2011).
47. Kabsch, W. XDS. *Acta Crystallogr. D* **66**, 125–132 (2010).
48. Evans, P. R. & Murshudov, G. N. How good are my data and what is the resolution? *Acta Crystallogr. D* **69**, 1204–1214 (2013).
49. Bricogne, G. *et al.* BUSTER version 2.11.5 <http://www.globalphasing.com> (2014).
50. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
51. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
52. The PyMOL Molecular Graphics System, Version 1.7.2.3 Schrödinger, LLC <https://www.pymol.org/> (2015).
53. Lu, X.-J. & Olson, W. K. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.* **31**, 5108–5121 (2003).
54. Nawrocki, E. P. *et al.* Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.* **43**, D130–D137 (2015).
55. Vaguine, A. A., Richelle, J. & Wodak, S. J. SFCHECK: a unified set of procedure for evaluating the quality of macromolecular structure-factor data and their agreement with atomic model. *Acta Crystallogr. D* **55**, 191–205 (1999).



Extended Data Figure 1 | Enzymatic steps responsible for riboflavin, FMN, and FAD biosynthesis in *E. coli*. **a**, Enzyme names are shown above reaction arrows. RibA is a GTP cyclohydrolase II, RibB is a (3S)-3,4-dihydroxy-2-butanone 4-phosphate synthase. RibDG is a bifunctional enzyme encoding 2,5-diamino-6-ribosylamino-4(3H)-pyrimidinone 5'-phosphate deaminase and 5-amino-6-ribosylamino-2,4(1H,3H)-pyrimidinedione 5'-phosphate reductase activity. RibH is a lumazine synthase. RibE is a riboflavin synthase. RibFC is another bifunctional enzyme encoding flaviokinase and FAD synthetase activity. Note, one molecule of GTP and two molecules of ribulose 5-phosphate are required to make one molecule of riboflavin. I, 2,5-diamino-6-ribosylamino-4(3H)-pyrimidinone 5' phosphate; II, 5-amino-6-ribosylamino-2,4(1H,3H)-pyrimidinedione 5'-phosphate; III, 5-amino-6-ribitylamino-2,4(1H,3H)-pyrimidinedione 5'-phosphate; IV, 5-amino-6-ribitylamino-2,4(1H,3H)-pyrimidinedione; V, (3S)-3,4-dihydroxy-2-butanone 4-phosphate; VI, 6,6-dimethyl-8-D-ribityllumazine. Note two molecules of VI dismutate (step 7) to give IV and riboflavin. Note also that the phosphatase converting III to IV (step 4) is not known. Adapted from Pedrolli *et al.*²⁵. **b**, HPLC-based quantitative analysis of riboflavin, FMN, and FAD levels in *E. coli* strain MB5746 following genetic inactivation of riboflavin

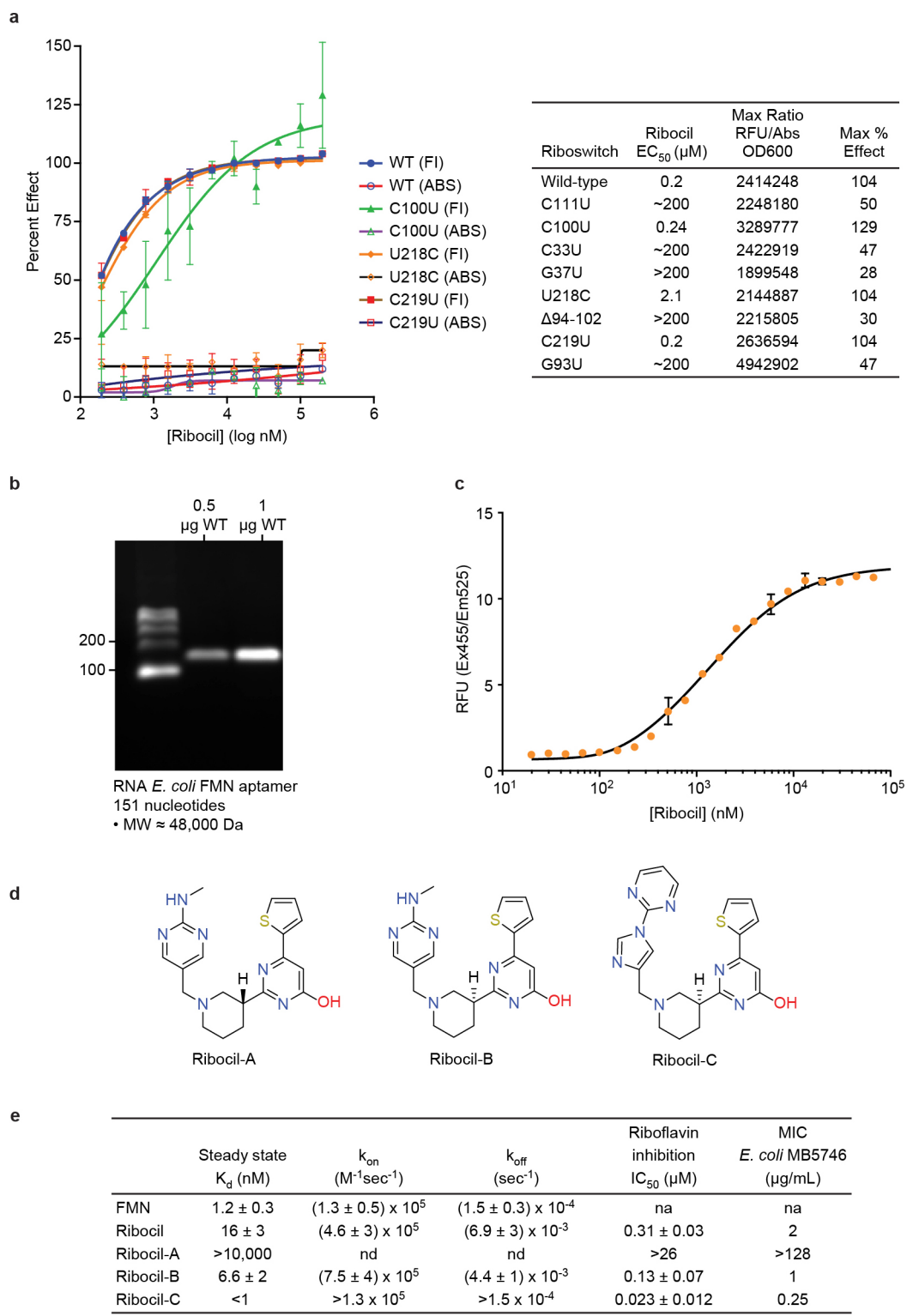
biosynthesis ($\Delta ribA$ or $\Delta ribB$) or ribocil drug treatment. Wild-type strain MB5746 and isogenic $\Delta ribA$ or $\Delta ribB$ strains were grown overnight in CAMHB media containing 10 μ M riboflavin, washed in CAMHB media lacking riboflavin, diluted (1:50) and grown for an additional 20 h before harvesting cells and HPLC analysis of cell lysates as described in the Methods. In parallel, overnight cultures of MB5746 grown in CAMHB were diluted (1:50) and treated with DMSO as a control or 20 μ M ribocil, grown for an additional 20 h and analysed as above (middle panel). Data are presented as the mean of two technical repeats and are representative of two independent experiments. Riboflavin, FMN and FAD depletion levels are listed as a percentage relative to wild-type controls (right panel). **c**, HPLC-based quantitative analysis of riboflavin levels in 19 independent *E. coli* ribocil^R mutant isolates versus the isogenic parent strain, MB5746. Overnight cultures of each strain were diluted 1:50 in CAMHB and treated with 10 μ M ribocil or mock treated (1% DMSO). After growth at 37 °C with agitation for 20 h, cell lysates were prepared and riboflavin levels were quantitated as described in the Methods. Data are the mean of two technical repeats (error bars indicate range) and is representative of two independent experiments.



Extended Data Figure 2 | *E. coli* FMN riboswitch-regulated reporter gene expression and ribocil-mediated inhibition and isolation of ribocil^R mutations mapping to *A. baumannii* and *P. aeruginosa* FMN riboswitches.

a, GFP reporter constructs under the control of the intact *E. coli* *ribB* promoter and FMN riboswitch (EcPro-EcFMN-GFP), or replaced with the *A. baumannii* FMN riboswitch (EcPro-AbFMN-GFP) or *P. aeruginosa* FMN riboswitch (EcPro-PaFMN-GFP) are maintained in *E. coli* ribocil^R mutant, M5 (Fig. 1e). Note, *E. coli* *ribB* upstream promoter sequence (EcPro) was fused to *A. baumannii* and *P. aeruginosa* FMN elements (AbFMN and PaFMN, respectively) to facilitate sufficient baseline expression in *E. coli* MB5746 host cells. **b**, Introduction of FMN riboswitch reporter plasmids into a ribocil^R mutant strain background (M5) enables ribocil dose-dependent inhibition of GFP expression without inhibiting cell growth. Dose-response curve for 16 technical repeats of ribocil-mediated inhibition of EcPro-EcFMN-GFP

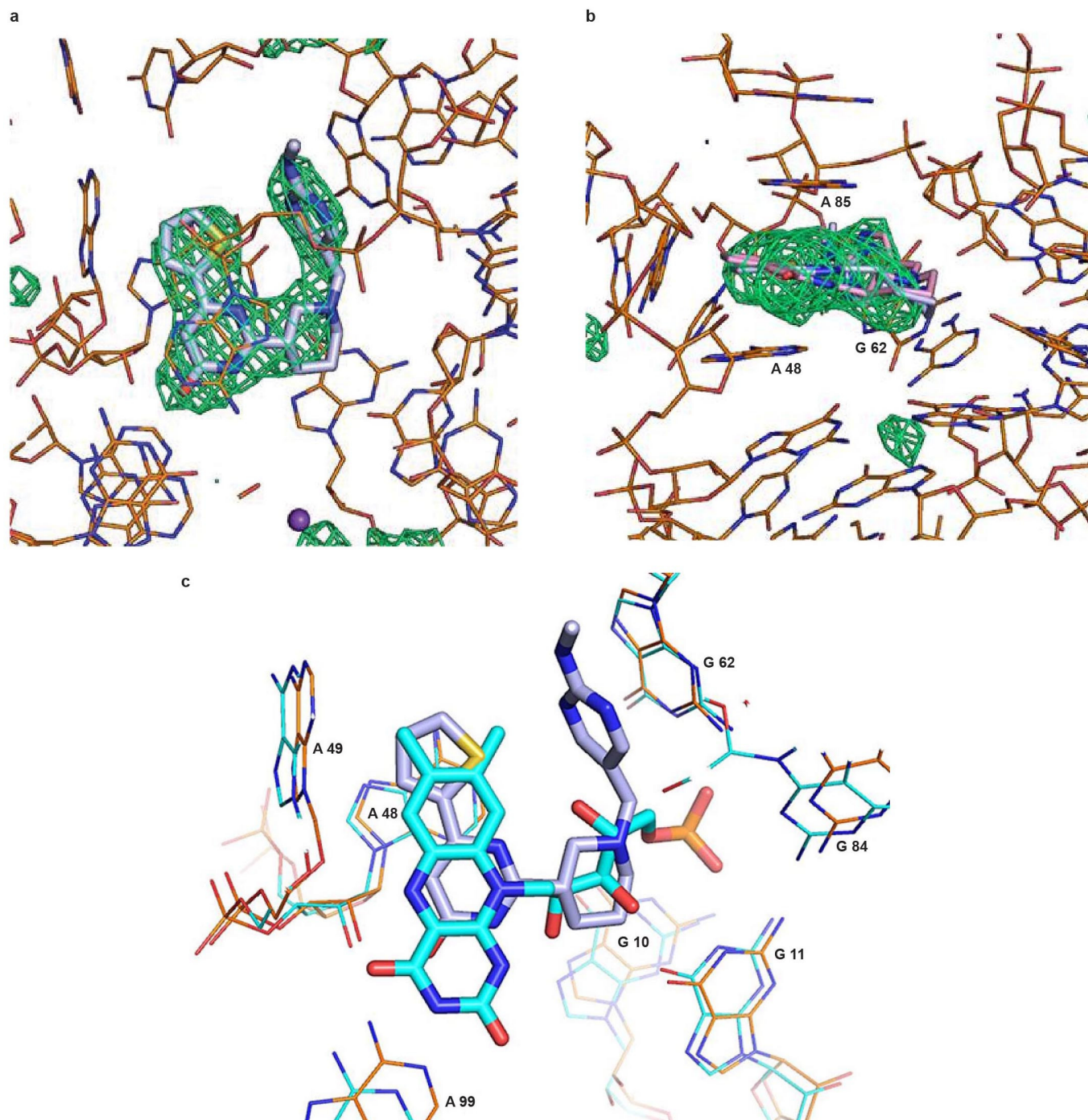
expression is shown. **c**, Tabulated EC₅₀ values (±s.d.) of ribocil required to inhibit EcPro-EcFMN-GFP, EcPro-AbFMN-GFP and EcPro-PaFMN-GFP expression. Data for EcPro-EcFMN-GFP are from eight independent experiments each with two technical repeats, whereas data for EcPro-AbFMN-GFP and EcPro-PaFMN-GFP are from two independent experiments each with four technical repeats. **d**, Schematic summary of *E. coli* recombined *ribB* locus, EcPro-AbFMN-*ribB*, in which the endogenous FMN riboswitch is replaced with the *A. baumannii* FMN riboswitch. Below, predicted secondary structure of the *A. baumannii* FMN riboswitch and ribocil^R mutations highlighted in red. **e**, Schematic summary of *E. coli* recombined *ribB* locus, EcPro-PaFMN-*ribB*, in which the *E. coli* FMN riboswitch is replaced by the *P. aeruginosa* FMN riboswitch. Below, predicted secondary structure of the *P. aeruginosa* FMN riboswitch and ribocil^R mutations highlighted in red.



Extended Data Figure 3 | Ribocil mutant analysis; Ribocil:FMN

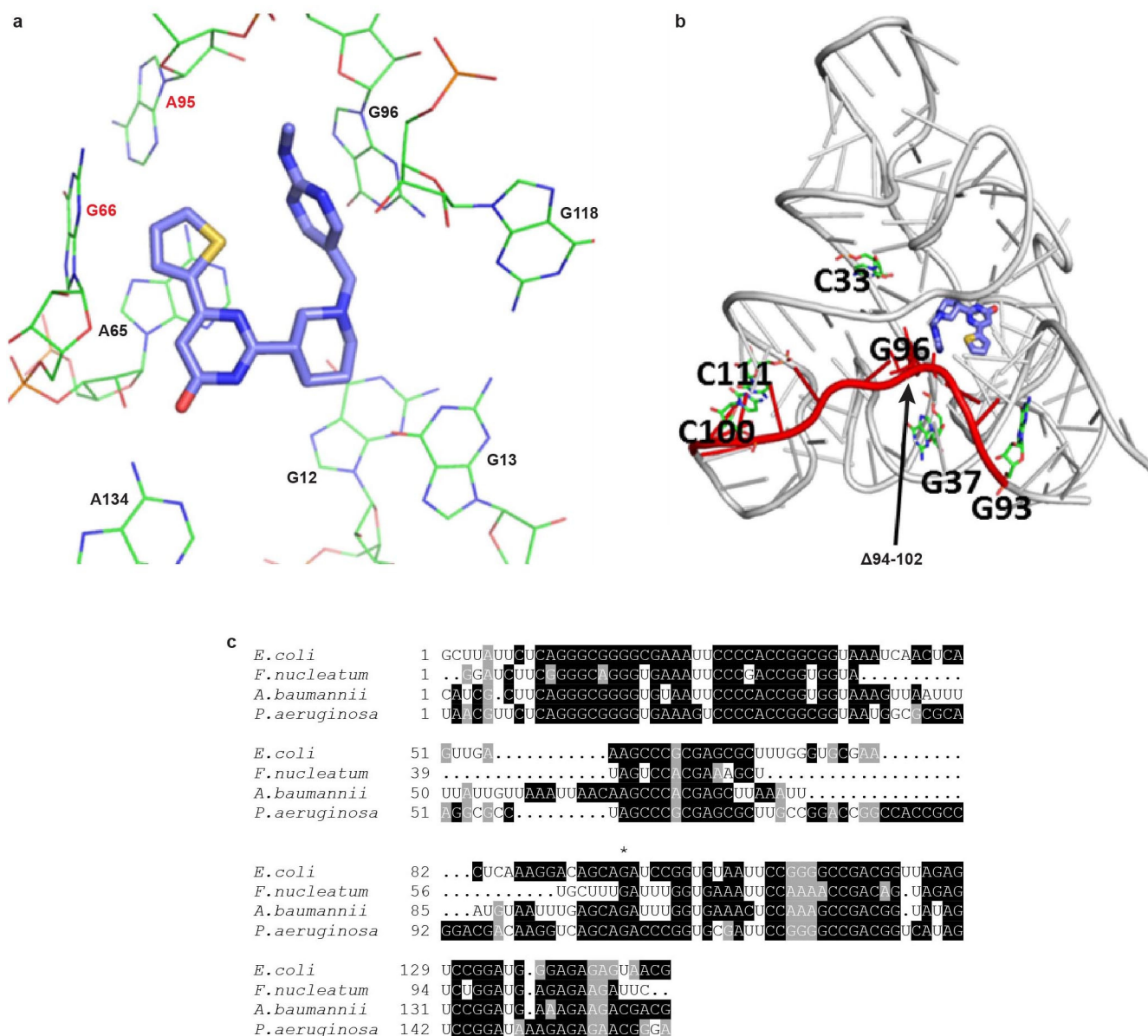
competition binding and Ribocil-analogue binding studies. **a**, Wild-type FMN and mutant constructs responsive to ribocil are displayed with a curve fit. Fluorescence was measured using 405 nm excitation and 510 nm emission and data represent the mean of two independent experiments (error bars indicate range). EC_{50} , the maximum cell-density-adjusted fluorescence signal observed with no ribocil addition, and the maximum per cent inhibition of fluorescence for each construct is listed in the accompanying table. **b**, Wild-type RNA aptamer samples were separated on 1.2% agarose gels and visualized by ethidium bromide staining. **c**, Binding affinity for ribocil. Binding affinity to the *E. coli* FMN riboswitch aptamer is determined from ribocil dose-dependent competition against a fixed concentration of FMN (60 nM) and a fixed concentration of *E. coli* FMN riboswitch aptamer (150 nM). Shown is a representative example of ribocil competition data. Mean affinity (\pm s.d.) from four independent experiments for ribocil, ribocil-A, ribocil-B and ribocil-C is reported in the table (panel **e** of this figure). **d**, Chemical structure of

ribocil-A, ribocil-B and ribocil-C. **e**, Steady-state K_d , binding kinetics, riboflavin biosynthesis inhibition and minimum inhibitory concentration (MIC) in *E. coli* MB5746 for ribocil-A, ribocil-B and ribocil-C. Binding kinetics were determined using a competition method employing a fixed amount of FMN (60 nM) and *E. coli* FMN riboswitch (48 nM). FMN binding affinity and kinetics are determined by fluorescence quenching experiments (see Methods). Given the tight-binding conditions, both FMN and ribocil affinity values represent an upper limit. As a result, the k_{off} values may also be an upper limit, and/or the k_{on} values may be a lower limit. Inhibition of riboflavin biosynthesis after treatment of *E. coli* MB5746 with ribocil, ribocil-A, ribocil-B and ribocil-C for 20 h as described in Methods. Data for ribocil are the mean (\pm s.d.) of three independent experiments, for ribocil-B the mean (\pm s.d.) of four independent experiments and for ribocil-C the mean (error bars indicate range) of two independent experiments. MIC of ribocil, ribocil-A, ribocil-B and ribocil-C against *E. coli* MB5746 determined by the broth microdilution method.



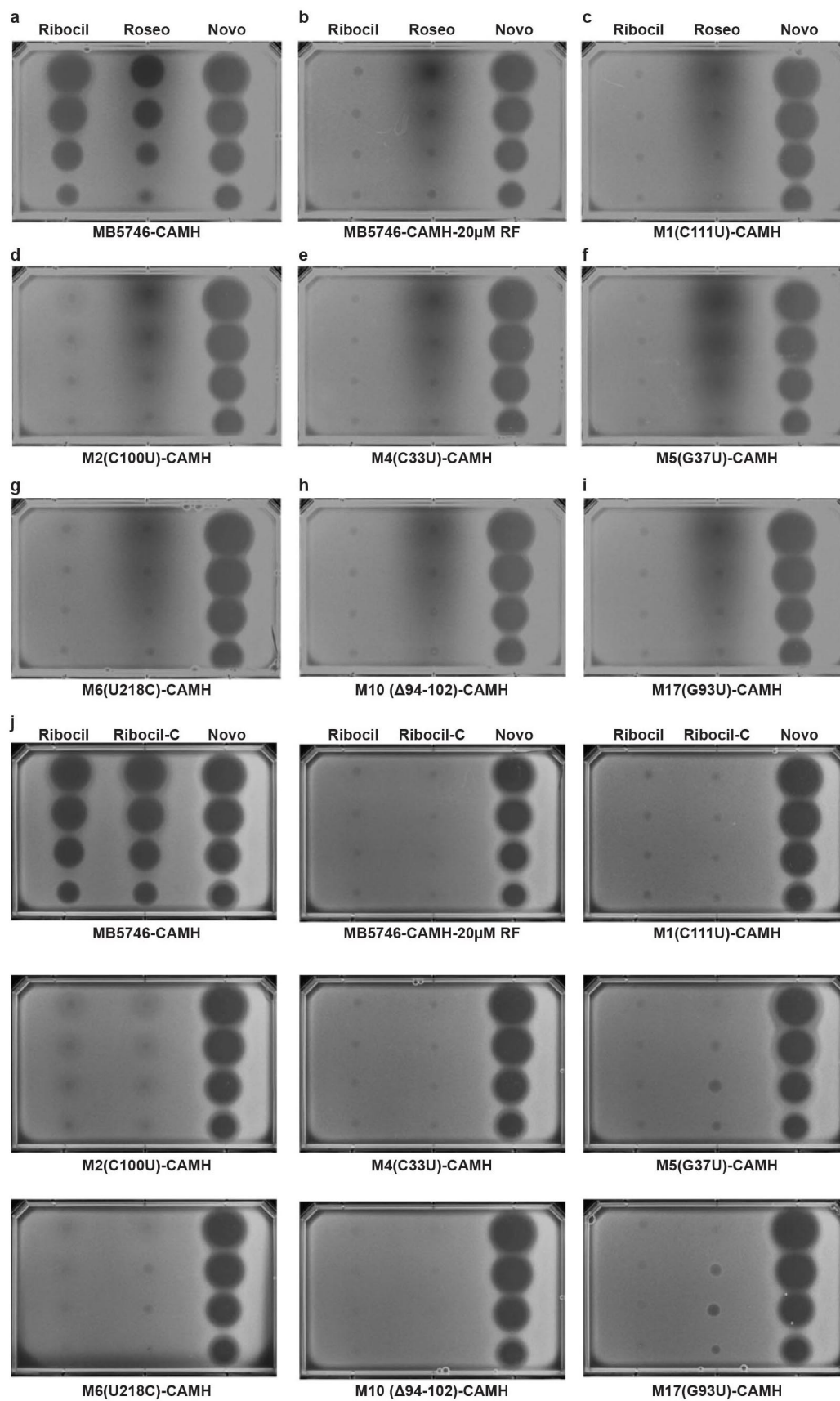
Extended Data Figure 4 | Electron density omit map of *F. nucleatum* impX FMN aptamer-ribocil co-crystal structure. **a**, The electron density difference map (see Methods for details regarding its calculation) is shown as a grid contoured at 3.0σ level. The refined structure of the ligand is shown as sticks, methyls in slate blue, the aptamer as lines with the methyl coloured in orange, cations as spheres coloured in purple. **b**, Same figure as **a** but rotated to bring the electron density around the planar 6-thiophenyl-pyrimidonyl horizontal and perpendicular to the figure plane. The best fit of the (S)-isomer which maintains the pyrimidinyl in its electron density, stacking against G62,

and inserts the 6-thiophenyl-pyrimidonyl between A48 and A85 is shown (see Methods). The 6-thiophenyl-pyrimidonyl is clearly slanted compared to the density and the planes of the A48 or A85 bases. **c**, Superposition of the X-ray co-structures of the *F. nucleatum* riboswitch aptamer with either FMN (PDB entry 3F2Q) or ribocil. After superposition using the phosphorus atoms of the bases in the immediate vicinity of the ligand, the RNA is represented as lines in cyan and orange for the co-structures with FMN and ribocil, respectively, and as sticks for the ligand, in cyan and slate blue, respectively. The number for the key bases interacting with either FMN or ribocil is indicated.



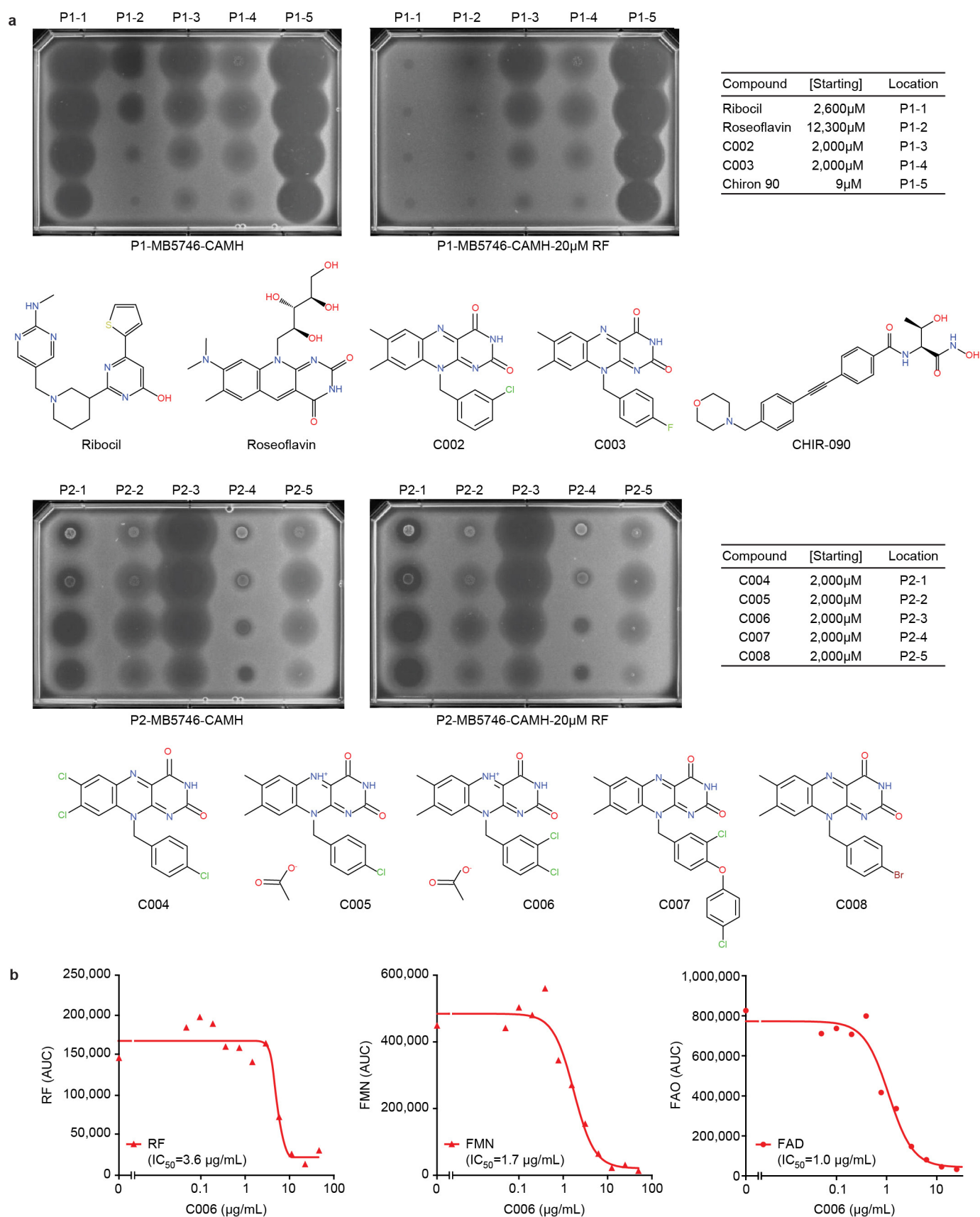
Extended Data Figure 5 | Homology model of the predicted ribocil binding site within the *E. coli* FMN aptamer. **a**, The homology model was constructed using program mutate_bases of the 3DNA package⁵³ using the *F. nucleatum* impX riboswitch aptamer X-ray structure as the template and the FMN aptamer alignment of *E. coli*, *F. nucleatum*, *P. aeruginosa* and *A. baumannii*. Only two bases differ (in red labels) compared to *F. nucleatum* FMN riboswitch, G66A and A95U. **b**, A full-length homology model of the *E. coli* FMN aptamer and mapping of ribocil^R mutations. Location of the ribocil^R mutants C33U, G37U, G93U, C100U, C111U and Δ94–102 are mapped on the *E. coli* homology model of the ribocil-bound *E. coli* FMN riboswitch aptamer. In the model, all nucleotide insertions in the *E. coli* sequence not found in

F. nucleatum were not modelled in the resulting crystal structure. There are 34 base changes among the 111 nucleotide modelled. A119 is removed for clarity. Nucleotides C33, G37, G93, C100 and C111 are coloured (green) and bases deleted in Δ94–102 are highlighted (red). G96 (red) which makes direct contact with ribocil (blue) in the wild-type aptamer is deleted in Δ94–102. **c**, Alignment of bacterial riboswitch sequences from RFAM⁵⁴. *E. coli* aptamer nucleotide G96, which is equivalent to nucleotide G62 in the *F. nucleatum* aptamer, is indicated with an asterisk. Black shading represents identical and grey is similar using default consensus settings with the BOXCHADE program (<http://sourceforge.net/projects/boxshade/>).



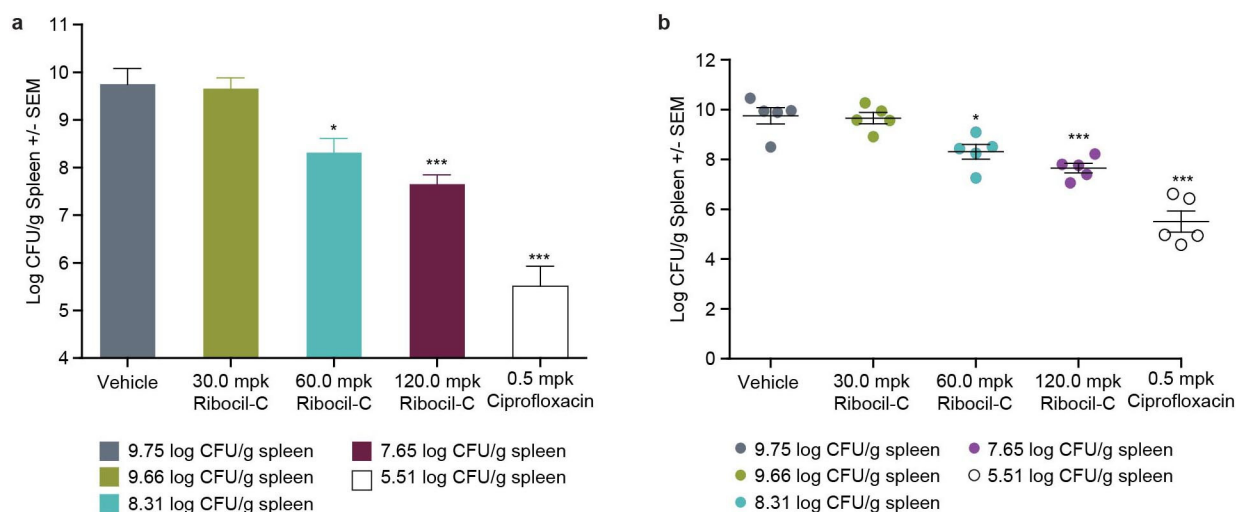
Extended Data Figure 6 | Ribocil, roseoflavin and ribocil-C cross-resistance to *E. coli* ribocil^R mutants. **a**, 5 μ l of ribocil (1.3 mM), roseoflavin (31 mM) and the negative control, novobiocin (2.5 mM), were spotted (twofold dilution series) on the surface of a CAMH plate (**a**) or CAMH plate plus 20 μ M riboflavin (**b**) and seeded with *E. coli* MB5746. **c–i**, Same as **a** but plates seeded with the indicated ribocil^R mutants. Note, whereas the growth-inhibiting activity of ribocil is completely suppressed by riboflavin, roseoflavin activity is only partially suppressed and ribocil^R mutants are cross-resistant to roseoflavin

and display similar level of suppression as riboflavin supplementation. The figure is representative of two independent experiments. **j**, Ribocil-C inhibits the FMN riboswitch and is cross-resistant to ribocil^R mutations. 5 μ l of ribocil (1.3 mM), ribocil-C (153 μ M) and the negative control, novobiocin (2.5 mM), were spotted (twofold dilution series) on the surface of CAMH plates seeded with the above described strains and/or riboflavin supplement as shown in **a–i**. The figure is representative of three independent experiments. RF, riboflavin.



Extended Data Figure 7 | Bioactivity of riboflavin analogues are not suppressed by riboflavin supplements. **a**, Left plates; 5 μl of ribocil (1.3 mM), roseoflavin (31 mM), C002 (2 mM), C003 (2 mM), C004 (2 mM), C005 (2 mM), C006 (2 mM), C007 (2 mM), C008 (2 mM), and the negative control, Chiron 90 (9 μM), were spotted (twofold dilution series) on the surface of a CAMH plate seeded with *E. coli* MB5746. Right plates; same as left plates

but supplemented with 20 μM riboflavin. Chemical structures are shown. The figure is representative of two independent experiments. **b**, Inhibition of flavin synthesis by C006. Flavin levels were determined by HPLC after C006 treatment of MB5746. All data are from a single 11-point dose titration and is representative of two independent experiments.



Extended Data Figure 8 | In vivo activity of ribocil-C in a murine systemic infection model of *E. coli*. **a**, DBA/2J mice were infected i.p. with *E. coli* strain MB5746 (5.0×10^5 CFU per mouse, a tenfold higher inoculum than that used in Fig. 4) and treated by subcutaneous injection with ribocil-C or ciprofloxacin at the indicated doses (mg kg^{-1}) three times over a 24 h infection period. Spleens were aseptically collected from five mice per group and the reduction of log[CFU per g spleen tissue] was calculated on the basis of bacterial burden in spleens of the vehicle-treated (10% DMSO) control group. Data represents the average CFU per g spleen \pm s.e.m. One-way ANOVA with

Dunnett's multiple comparison test demonstrates statistically significant ($*P < 0.05$, $***P < 0.001$) log[CFU per g spleen] reductions in the 60 and 120 mg kg^{-1} ribocil-C treatment groups and ciprofloxacin control. **b**, The CFU data from each mouse plotted as individual points, solid bar and error bars represent the average CFU per g spleen \pm s.e.m. One-way ANOVA with Dunnett's multiple comparison test demonstrates statistically significant ($*P < 0.05$, $***P < 0.001$) log[CFU per g spleen] reductions in the 60 and 120 mg kg^{-1} ribocil-C treatment groups and ciprofloxacin control.

Extended Data Table 1 | Frequency of resistance (FOR) determination and microbiological activity summary

a

MIC for Ribocil 2 µg/mL 8X MIC = 16 µg/mL 1x10 ⁻⁷ CFU/plate, 3 plates/condition		
FOR		
	CAMH Media	Minimal Media
MB5746	2.6x10 ⁻⁶	3.4x10 ⁻⁶
MB5746-Ab-RFN	3.3x10 ⁻⁸	<3.3x10 ⁻⁸
MB5746-Pa-RFN	6.4x10 ⁻⁷	ND

b

Compound	Highest test conc. (µg/mL)	<i>Escherichia coli</i>			<i>E. coli</i> MB5746/ <i>Pseudomonas aeruginosa</i> RFN ribB			<i>E. coli</i> MB5746/ <i>Acinetobacter baumannii</i> RFN ribB		
		MB5008 <i>lpxC</i> -	MB5747 <i>tolC::tn10</i>	MB5746 <i>lpxC</i> - <i>tolC::tn10</i>	Pa-WT	Pa-2 (-193 mut)	Pa-17 (-317 mut)	Ab-WT	Ab-14 (-57 mut)	Ab-17 (-9 mut)
Ribocil	64	16	2	2	16	>64	>64	2	>64	>64
Roseoflavin	128	>128	>128	>128	4	>128	>128	4	>128	>128
Ribocil-A	128	>128	>128	>128	>128	>128	>128	64	>128	>128
Ribocil-B	128	8	1	1	8	>128	>128	1	128	>128
Ribocil-C	64	2	0.25	0.25	>64	>64	>64	≤0.063	4	>64
Penicillin G	64	>64	>64	>64	>64	>64	>64	>64	>64	>64
Novobiocin	64	4	0.5	0.5	0.25	0.25	0.25	0.25	0.25	0.5
Chloramphenicol	64	1	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Nalidixic acid	64	1	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5

c

Compound	Highest test conc. (µg/mL)	<i>Pseudomonas aeruginosa</i>		<i>Acinetobacter baumannii</i>	
		ATCC35151 Z61	MB5890 Multi-efflux	ATCC12457 Δ <i>lpxA</i>	ATCC19606 Δ <i>lpxC</i>
Ribocil	128	>128	>128	>128	>128
Ribocil-C	64	>64	>64	>64	>64
Roseoflavin	128	>128	>128	>128	>128
Penicillin G	64	≤0.063	>64	>64	32
Novobiocin	64	0.5	64	8	4
Chloramphenicol	64	0.5	0.5	8	8
Nalidixic acid	64	1	0.5	0.5	0.25

d

Compound	Highest test conc. (µg/mL)	<i>Saccharomyces cerevisiae</i>		<i>Candida albicans</i>		
		S288c	Sc Δ <i>pdr5</i>	CAF2-1 <i>CDR1/CDR1</i> <i>CDR2/CDR2</i>	DSY448 Δ <i>cdri</i> /Δ <i>cdri</i> <i>CDR2/CDR2</i>	DSY654 Δ <i>cdri</i> /Δ <i>cdri</i> Δ <i>cdri</i> /Δ <i>cdri</i>
Ribocil	32	>32	>32	>32	>32	>32
Ribocil-C	32	>32	>32	>32	>32	>32
Roseoflavin	32	>32	>32	>32	>32	>32
Fluconazole	16	4	0.5	>16 (0.5)	1 (0.25)	1 (0.25)
Amphotericin B	16	1	1	1	1	1
Caspofungin	4	0.5 (0.25)	0.5 (0.25)	0.125	0.125	0.125

a, Frequency of resistance to ribocil by *E. coli* strain MB5746 carrying *A. baumannii*, *P. aeruginosa* or native RFN. Testing was performed with both Mueller Hinton (CAMH) and M9 (Minimal Media) broth.

b, Antibacterial activity of ribocil analogues and standard controls tested against: (1) efflux- and permeability-deficient *E. coli* MB5746 and two progenitor strains; (2) *E. coli* strain MB5746 carrying wild-type and mutant alleles of *A. baumannii* and *P. aeruginosa* RFN. Minimum inhibitory concentrations (MICs) were determined by CLSI standard broth microdilution. c, Antibacterial activity of ribocil analogues and standard controls tested against sensitized *P. aeruginosa* and *A. baumannii* isolates. MICs were determined by CLSI standard broth microdilution. d, Antifungal activity of ribocil analogues and standard controls tested against wild-type and efflux-deficient *Saccharomyces cerevisiae* and *Candida albicans*. MICs were determined by CLSI standard broth microdilution. All MIC values are expressed as µg ml⁻¹.

Extended Data Table 2 | X-ray crystal data collection and refinement statistics

FMN riboswitch-ribocil	
Data collection	
Space group	P 3 ₁ 2 1
Cell dimensions	
<i>a</i> , <i>b</i> , <i>c</i> (Å)	71.1, 71.1, 138.5
α , β , γ (°)	90.0, 90.0, 120.0
Resolution (Å)	2.93 (3.27-2.93)*
<i>R</i> _{merge}	3.8% (39.0%)
<i>I</i> / σ <i>I</i>	36.2 (6.2)
Completeness (%)	99.9% (100%)
Redundancy	9.5(9.8)
Refinement	
Resolution (Å)	35.6 – 2.93
No. reflections	Work 8,737 free 442
<i>R</i> _{work} / <i>R</i> _{free}	21.3%/23.4%
No. non-H atoms	
RNA	2,340
Ligand/ion	27/8
Water	0
B-factors	
RNA	103.3
Ligand/ion	85.8/113.7
Water	N/A
R.m.s deviations	
Bond lengths (Å)	0.0098
Bond angles (°)	0.74
Coord. error from Luzzati plot (Å)	0.452
Molprobability score	2.33 (97th percentile)

A full data set was collected with one crystal. Coordinate error estimated from Luzzati plot using SFCHECK³⁵.

*Highest resolution shell is shown in parenthesis.

Abundant molecular oxygen in the coma of comet 67P/Churyumov–Gerasimenko

A. Bieler^{1,2}, K. Altwegg^{2,3}, H. Balsiger², A. Bar-Nun⁴, J.-J. Berthelier⁵, P. Bochslers², C. Brioso⁶, U. Calmonte², M. Combi¹, J. De Keyser⁷, E. F. van Dishoeck⁸, B. Fiethe⁹, S. A. Fuselier¹⁰, S. Gasc², T. I. Gombosi¹, K. C. Hansen¹, M. Hässig^{2,10}, A. Jäckel², E. Kopp², A. Korth¹¹, L. Le Roy³, U. Mall¹¹, R. Maggiolo⁷, B. Marty¹², O. Mousis¹³, T. Owen¹⁴, H. Rème^{15,16}, M. Rubin², T. Sémon², C.-Y. Tzou², J. H. Waite¹⁰, C. Walsh⁸ & P. Würz^{2,3}

The composition of the neutral gas comas of most comets is dominated by H₂O, CO and CO₂, typically comprising as much as 95 per cent of the total gas density¹. In addition, cometary comas have been found to contain a rich array of other molecules, including sulfuric compounds and complex hydrocarbons. Molecular oxygen (O₂), however, despite its detection on other icy bodies such as the moons of Jupiter and Saturn^{2,3}, has remained undetected in cometary comas. Here we report *in situ* measurement of O₂ in the coma of comet 67P/Churyumov–Gerasimenko, with local abundances ranging from one per cent to ten per cent relative to H₂O and with a mean value of 3.80 ± 0.85 per cent. Our observations indicate that the O₂/H₂O ratio is isotropic in the coma and does not change systematically with heliocentric distance. This suggests that primordial O₂ was incorporated into the nucleus during the comet's formation, which is unexpected given the low upper limits from remote sensing observations⁴. Current Solar System formation models do not predict conditions that would allow this to occur.

Measurements of the coma of 67P/Churyumov–Gerasimenko (hereafter 67P) were made between September 2014 and March 2015 with the ROSINA-DFMS mass spectrometer⁵ on board the Rosetta spacecraft. For the present study, we analysed 3,193 mass spectra taken in this time period. Because of the high mass resolving power and sensitivity of ROSINA-DFMS, it was possible to differentiate unambiguously between the three main species present in the narrow mass range centred at mass-to-charge (*m/z*) ratio 32 Da/e, namely, molecular oxygen (O₂), atomic sulfur (S) and methanol (CH₃OH); such differentiation has not been achieved by previous *in situ* or remote sensing measurements at comets. Figure 1 shows several DFMS measurements centred at the O₂ peak. The green and orange lines show data taken before the close encounter with 67P. Only minor signatures from the tenuous neutral gas atmosphere of the Rosetta spacecraft can be identified, and even after long thruster firing manoeuvres, which use N₂O₄ as an oxidizer, the contamination of the O₂ signal remains small (see the green line in Fig. 1). Measurements taken while orbiting 67P, shown as the light blue, dark blue and purple lines in Fig. 1, show a clear increase of the O₂ signal, indicative of cometary O₂. These three measurements were taken at decreasing distances (*r*) from the comet nucleus, and follow the predicted $1/r^2$ signal dependence that is expected for conserved cometary species, giving further confidence in our detection.

As previously reported^{6,7}, the local number densities in the coma vary spatially and temporally^{6,7} for different compounds. Figure 2 shows

correlation plots of H₂O with O₂, CO and N₂. Red symbols represent data from 17 to 23 October 2014, for which previously published data on N₂ are available. During this time the spacecraft was in closed orbits at 10 km from the nucleus centre. Black symbols are data from 1 September 2014 to 31 March 2015. For both periods, O₂ clearly shows the strongest correlation with H₂O. While CO shows a high correlation with H₂O from 17 to 23 October 2014, the correlation for the whole data set is fairly low. N₂ shows the weakest correlation with H₂O of all three species. The strong correlation between H₂O and O₂, with a Pearson correlation coefficient of 0.88 (and even 0.97 for the October data), indicates that they are of similar origin in the nucleus and that their release mechanisms are linked, in contrast to CO and

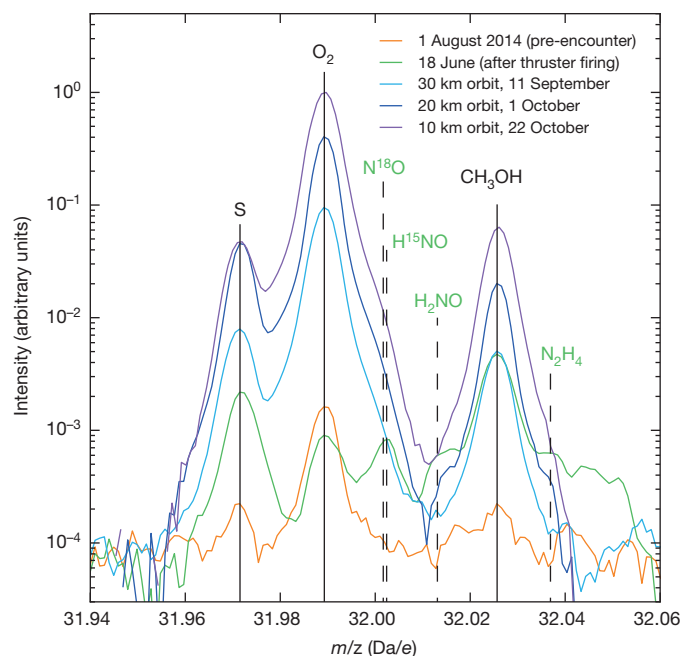


Figure 1 | DFMS mass spectra around 32 Da/e normalized to the spectrum with the largest signal. The black labels indicate the three major species found in the coma of 67P at 32 Da/e. The green labels and green line identify contamination peaks from thruster firings, showing that their contributions to the O₂ peak are very low. The light blue, dark blue and purple lines represent measurements taken at different distances from the comet nucleus.

¹Department of Climate and Space Science and Engineering, University of Michigan, 2455 Hayward Street, Ann Arbor, Michigan 48109, USA. ²Physikalisches Institut, University of Bern, Sidlerstrasse 5, CH-3012 Bern, Switzerland. ³Center for Space and Habitability, University of Bern, Sidlerstrasse 5, CH-3012 Bern, Switzerland. ⁴Department of Geosciences, Tel-Aviv University, Ramat-Aviv, 6997801 Tel-Aviv, Israel. ⁵LATMOS/IPSL-CNRS-UPMC-UVSQ, 4 Avenue de Neptune, F-94100 Saint-Maur, France. ⁶Laboratoire de Physique et Chimie de l'Environnement et de l'Espace (LPC2E), UMR 6115 CNRS – Université d'Orléans, 45071 Orléans, France. ⁷Belgian Institute for Space Aeronomy, BIRA-IASB, Ringlaan 3, B-1180 Brussels, Belgium. ⁸Leiden Observatory, Leiden University, PO Box 9513, 2300 RA Leiden, The Netherlands. ⁹Institute of Computer and Network Engineering (IDA), TU Braunschweig, Hans-Sommer-Straße 66, D-38106 Braunschweig, Germany. ¹⁰Space Science and Engineering Division, Southwest Research Institute, 6220 Culebra Road, San Antonio, Texas 78228, USA. ¹¹Max-Planck-Institut für Sonnensystemforschung, Justus-von-Liebig-Weg 3, 37077 Göttingen, Germany. ¹²Centre de Recherches Pétrographiques et Géochimiques, CRPG-CNRS, Université de Lorraine, 15 rue Notre Dame des Pauvres, BP 20, 54501 Vandœuvre lès Nancy, France. ¹³Aix Marseille Université, CNRS, LAM (Laboratoire d'Astrophysique de Marseille) UMR 7326, 13388 Marseille, France. ¹⁴Institute for Astronomy, University of Hawaii, Honolulu, Hawaii 96822, USA. ¹⁵Université de Toulouse-UPS-OMP-IRAP, 31400 Toulouse, France. ¹⁶CNRS-IRAP, 9 avenue du Colonel Roche, BP 44346, F-31028 Toulouse Cedex 4, France.

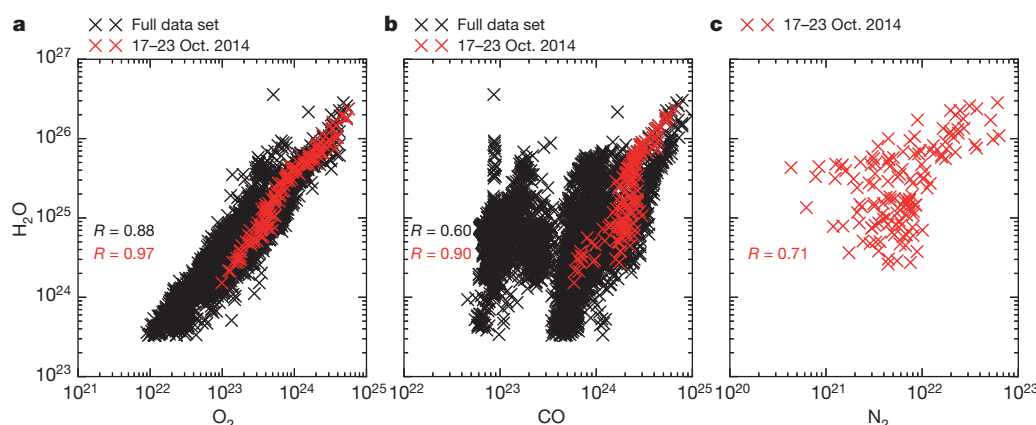


Figure 2 | Correlation between H₂O and O₂, CO and N₂. **a**, H₂O and O₂; **b**, H₂O and CO; **c**, H₂O and N₂. All three panels share a common y axis. Numbers on x and y axes are proportional to number density but in arbitrary units. Red crosses mark a subset of data for which N₂ data are also available. Panel **a** shows the strong correlation between H₂O and O₂, which is

observed for all data. In contrast, the correlation of CO with H₂O (**b**) varies over time, which leads to a low overall correlation between those two species. N₂ has the lowest correlation with H₂O of the compared species for the October data (**c**).

N₂ which have a similar volatility but do not show a strong correlation with H₂O (see Fig. 2 for correlation coefficient values). The O₂/H₂O ratio decreases for high H₂O abundances, which might be caused by surface water ice produced by a cyclic sublimation–condensation process⁸, although the total amount of surface ice is limited⁹.

A plausible mechanism for the strong O₂/H₂O correlation would be the production of O₂ by radiolysis or photolysis of water ice. Here we follow the convention that photolysis refers to ultraviolet photons breaking bonds, whereas radiolysis refers to more energetic photons or fast electrons and ions depositing energy into the ice and ionizing molecules. Creation of sputtered O₂ by radiolysis has been demonstrated in laboratory experiments¹⁰ and is observed for the icy moons of Jupiter—Europa, Ganymede and Callisto^{11–13}—as well as for the rings of Saturn³. Comets are subject to radiolysis over various timescales: (a) over billions of years, while they reside in the Kuiper belt; (b) over the period of a few years once they enter the inner Solar System; and (c) on very short timescales, as for the present radiolysis. In the Kuiper belt, the skin depth for producing O₂ is in the range of metres, although the produced O₂ may diffuse deeper into the porous nucleus. Once a comet begins its residence in the inner Solar System, it loses its surface material to a depth of several metres during each orbit around the Sun, therefore we can safely assume that no O₂ from radiolysis in the Kuiper belt phase remains in 67P at the percentage level. Radiolysis and photolysis by solar wind and ultraviolet radiation in the inner Solar System only affect the top few micrometres of the cometary

surface. Taking account of 67P's continuous mass loss through outgassing, we estimate the actively outgassing surface areas to be lost to a depth of several centimetres over the time from August 2014 to March 2015. If recent production by radiolysis or photolysis (only affecting the top few micrometres) were the source of the measured O₂, our data would show a continuous decrease of the O₂/H₂O ratio over the examined time period as the active surface continues to be shed over that time. Apart from the variations related to H₂O abundance, Fig. 3 shows that we do not observe a systematic change in the O₂/H₂O ratio over several months. Instantaneous creation of the measured O₂ by radiolysis or photolysis seems, overall, unlikely, and would lead to variable O₂ ratios due to different illumination conditions. Given that radiolysis and photolysis, on any of the discussed timescales, do not seem to be plausible production mechanisms, the preferred explanation of our observations is the incorporation of primordial O₂ into the cometary nucleus.

Despite great efforts by remote sensing campaigns, information on primordial O₂ is still limited. Solid O₂ has not yet been detected in interstellar ices, and upper limits for the O₂/H₂O ice ratios of <0.5 and for O₂/CO ratios of <1 are in agreement with our findings, but such high upper limits do not provide useful constraints^{14,15}. Gaseous O₂ has only been detected in two interstellar clouds so far^{4,16,17}, and is generally known to have surprisingly low abundances⁴. Reports of very low upper limits for O₂ in a protostellar envelope suggest the material infalling to the accretion disk is very poor in molecular oxygen¹⁸. This has been ascribed

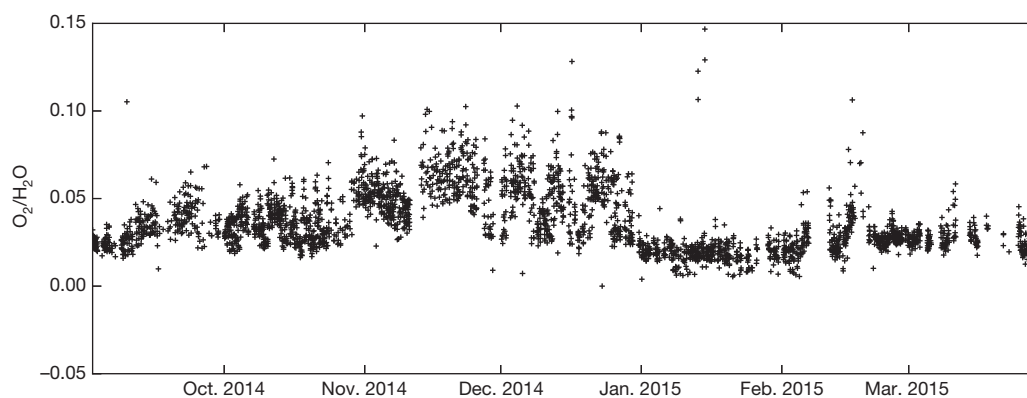


Figure 3 | O₂/H₂O ratio over several months. There seems to be no systematic increase or decrease of the O₂/H₂O ratio. The variances happen on very short timescales and can be explained by the decrease of the O₂ ratio for high H₂O abundances. It is not fully understood if the higher variability

of the O₂ ratio from October to the end of December 2014 can be attributed to orbital changes of the spacecraft or to physical changes of the cometary nucleus.

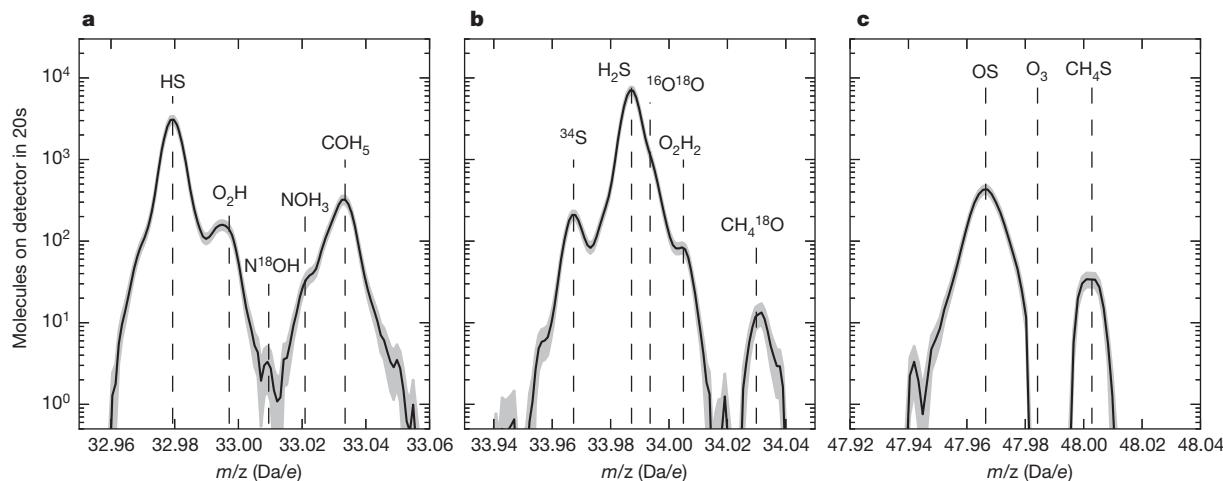


Figure 4 | DFMS spectra for some of the common products of radiolysis of water ice. **a–c**, Products are O_2H (seen in **a**), $^{16}\text{O}^{18}\text{O}$ and O_2H_2 (seen in **b**) and O_3 (not seen in **c**). These data were recorded on 20 October 2014 at around 01:00 UTC. With the exception of O_3 (**c**), all the previously mentioned

species are measured and can clearly be identified in the mass spectra of DFMS. The grey area consists of the statistical error N/\sqrt{N} and a 10% uncertainty from the individual pixel gains on the detector (here N is the number of molecules on the detector).

to the high volatility and reactivity of O_2 , in particular the rapid transformation of O and O_2 to H_2O ice on cold grains¹⁹.

However, in one of the two interstellar clouds where O_2 has been detected (the ρ Ophiuchi dense core), the chemically related species HO_2 and H_2O_2 have been measured^{20,21} with gaseous abundance ratios of $\text{HO}_2/\text{O}_2 \approx \text{H}_2\text{O}_2/\text{O}_2 \approx 0.6 \times 10^{-3}$. Interestingly, the abundance ratios determined by DFMS for the coma of 67P are very close to these interstellar values: $\text{HO}_2/\text{O}_2 = (1.9 \pm 0.3) \times 10^{-3}$ and $\text{H}_2\text{O}_2/\text{O}_2 = (0.6 \pm 0.07) \times 10^{-3}$ (see Fig. 4). If these gas-phase abundance ratios reflect those in the cometary ice, this would support the existence of primordial O_2 . The ρ Ophiuchi A core has been suggested to have experienced slightly higher temperatures of around 20–30 K over its lifetime (which is also typical of estimates for the comet-forming conditions in the outer early solar nebula), compared to ~ 10 K for most other dense interstellar clouds^{21,22}. If higher temperatures are indeed needed to produce significant amounts of O_2 , this would indicate that our Solar System was formed from an unusually warm molecular cloud, consistent with the low abundance of N_2 in 67P (ref. 23).

One aspect of the present results that remains unexplained is the high value (a few per cent) of the $\text{O}_2/\text{H}_2\text{O}$ ratio in 67P. Models of gas-grain chemistry in molecular clouds predict $\text{O}_2/\text{H}_2\text{O}$ ratios at least an order of magnitude lower¹⁹. They also over-predict ozone (O_3): we found no evidence for the presence of ozone (see Fig. 4), with an upper limit of 1×10^{-6} relative to water.

An alternative explanation for the presence of O_2 is the incorporation of gaseous O_2 into water ice in the protosolar nebula, and chemical models of protoplanetary disks do in fact show high abundances of gaseous O_2 in the comet forming zone²⁴. Rapid cooling from >100 K to less than 30 K would then be needed to form amorphous water ice with trapped O_2 on dust grains. This could happen when young disks experience increased heating due to accretion bursts onto the star, followed by a rapid drop in temperature as soon as the burst is over. These O_2 rich grains then need to be accreted into larger bodies before further chemical modification occurs.

Finally, we discuss radiolysis of icy grains before accretion. When produced by radiolysis in water ice, O_2 can remain trapped in voids, while hydrogen can diffuse out¹⁰. This prevents the hydrogenation of O_2 , which is otherwise a dominant reaction for the destruction of molecular oxygen, and could lead to increased and stable levels of O_2 in the solid ice²⁵. Incorporation of such icy grains into the comet nucleus would explain the observed strong correlation with H_2O , in contrast to N_2 which is trapped from the gas phase and shows a lower correlation with water. However, O_3 (resulting from O_2 radiolysis) has been

reported to be trapped in Ganymede's surface²⁵, and at such a concentration would just be detectable by DFMS for O_2 levels at a few per cent of H_2O , but no O_3 could be detected. A further consequence would be that these icy grains have been incorporated into the comet mostly unaltered, a process that is much debated, but which has recently been proposed again²⁶ and that would also be in accordance with the measured high D/H ratio in 67P (ref. 27).

We note that our findings do not significantly affect our understanding of the global distribution of elemental O in the interstellar medium, as O_2 in ice with an abundance of a few per cent relative to H_2O accounts only for a small fraction of the total oxygen inventory.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 29 June; accepted 18 August 2015.

1. Bockelée-Morvan, D., Mumma, M. J. & Weaver, H. A. in *Comets II* (eds Festou, M., Keller, U. H. & Weaver, H. A.) 391–423 (Univ. Arizona Press, 2004).
2. Hall, D. T., Strobel, D. F., Feldman, P. D., McGarth, M. A. & Weaver, H. A. Detection of an oxygen atmosphere on Jupiter's moon Europa. *Nature* **373**, 677–679 (1995).
3. Johnson, R. E. *et al.* Production, ionization and redistribution of O_2 in Saturn's ring atmosphere. *Icarus* **180**, 393–402 (2006).
4. Goldsmith, P. F. *et al.* Herschel measurements of molecular oxygen in Orion. *Astrophys. J.* **737**, 96 (2011).
5. Balsiger, H. *et al.* ROSINA — ROSETTA orbiter spectrometer for ion and neutral analysis. *Space Sci. Rev.* **128**, 745–801 (2007).
6. Hässig, M. *et al.* Time variability and heterogeneity in the coma of 67P/Churyumov-Gerasimenko. *Science* **347**, aaa0276 (2015).
7. Luspai-Kuti, A. *et al.* Composition-dependent outgassing of comet 67P/Churyumov-Gerasimenko from ROSINA/DFMS — implications for nucleus heterogeneity? *Astron. Astrophys.* <http://dx.doi.org/10.1051/0004-6361/201526205> (2015).
8. De Sanctis, M. C. *et al.* The diurnal cycle of water ice on comet 67P/Churyumov-Gerasimenko. *Nature* **525**, 500–503 (2015).
9. Capaccioni, F. *et al.* The organic-rich surface of comet 67P/Churyumov-Gerasimenko as seen by VIRTIS/Rosetta. *Science* **347**, aaa0628 (2015).
10. Brown, W. L. *et al.* Erosion and molecular formation in condensed gas films by electronic energy loss of fast ions. *Nucl. Instrum. Methods* **198**, 1–8 (1982).
11. Carlson, R. W. *et al.* Hydrogen peroxide on the surface of Europa. *Science* **283**, 2062–2064 (1999).
12. Spencer, J. R., Calvin, W. M. & Person, M. J. Charge-coupled device spectra of the Galilean satellites: molecular oxygen on Ganymede. *J. Geophys. Res.* **100**, 19049–19056 (1995).
13. Spencer, J. R. & Calvin, W. M. Condensed O_2 on Europa and Callisto. *Astron. J.* **124**, 3400–3403 (2002).
14. Vandenbussche, B. *et al.* Constraints on the abundance of solid O_2 in dense clouds from ISO-SWS and ground-based observations. *Astron. Astrophys.* **346**, L57–L60 (1999).
15. Pontoppidan, K. *et al.* A 3–5 μm VLT spectroscopic survey of embedded young low mass stars I. Structure of the CO ice. *Astron. Astrophys.* **408**, 981–1007 (2003).

16. Liseau, R. *et al.* Multi-line detection of O₂ toward ρ Ophiuchi A. *Astron. Astrophys.* **541**, A73 (2012).
17. Larsson, B. *et al.* Molecular oxygen in the Ophiuchi cloud. *Astron. Astrophys.* **466**, 999–1003 (2007).
18. Yildiz, U. A. *et al.* Deep observations of O₂ toward a low-mass protostar with Herschel-HIFI. *Astron. Astrophys.* **558**, A58 (2013).
19. Taquet, V., Ceccarelli, C. & Kahane, C. Multilayer modeling of porous grain surface chemistry. I. The GRAINOBLE model. *Astron. Astrophys.* **538**, A42 (2012).
20. Bergman, P. *et al.* Detection of interstellar hydrogen peroxide. *Astron. Astrophys.* **531**, L8 (2011).
21. Parise, B., Bergman, P. & Du, F. Detection of the hydroperoxyl radical HO₂ toward ρ Ophiuchi A. Additional constraints on the water chemical network. *Astron. Astrophys.* **541**, L11 (2012).
22. Du, F., Parise, B. & Bergman, P. Production of interstellar hydrogen peroxide (H₂O₂) on the surface of dust grains. *Astron. Astrophys.* **538**, A91 (2012).
23. Rubin, M. *et al.* Molecular nitrogen in comet 67P/Churyumov-Gerasimenko indicates a low formation temperature. *Science* **348**, 232–235 (2015).
24. Walsh, C., Nomura, H. & van Dishoeck, E. The molecular composition of the planet-forming regions of protoplanetary disks across the luminosity range. *Astron. Astrophys.* **582**, A88 (2015).
25. Johnson, R. E. & Jessor, W. A. O₂/O₃ microatmospheres in the surface of Ganymede. *Astrophys. J.* **480**, L79–L82 (1997).
26. Cleaves, L. I. *et al.* The ancient heritage of water ice in the solar system. *Science* **345**, 1590–1593 (2014).
27. Altwegg, K. *et al.* 67P/Churyumov-Gerasimenko, a Jupiter family comet with a high D/H ratio. *Science* **347**, <http://dx.doi.org/10.1126/science.1261952> (2015).

Acknowledgements Work at the University of Michigan was funded by NASA contract JPL-1266313. Work at the University of Bern was funded by the State of Bern, the Swiss

National Science Foundation and the European Space Agency PRODEX Program. Work at Max-Planck-Institut für Sonnensystemforschung was funded by the Max-Planck Society and BMWI contract 50QP1302. Work at Southwest Research Institute was supported by subcontract 1496541 from the Jet Propulsion Laboratory. Work at BIRA-IASB was supported by the Belgian Science Policy Office via PRODEX/ROSINA PEA 90020. This work was carried out thanks to the support of the A*MIDEX project (no. ANR-11-IDEX-0001-02) funded by the 'Investissements d'Avenir' French Government programme, managed by the French National Research Agency (ANR). This work was supported by CNES grants at IRAP, LATMOS, LPC2E, UTINAM, CRPG, and by the European Research Council (grant no. 267255 to B.M.). A.B.-N. thanks the Ministry of Science and the Israel Space agency. Work by J.H.W. at Southwest Research Institute was funded by NASA JPL subcontract NAS703001TONMO710889. E.F.v.D. and C.W. are supported by A-ERC grant 291141 CHEMPLAN and an NWO Veni award. We acknowledge here the work of the whole ESA Rosetta team.

Author Contributions A.B. performed data reduction, analysis and wrote the paper; K.A. initialized and edited the paper and contributed to data interpretation; C.B., U.C., M.C., T.J.G., K.C.H., S.G., M.H., A.J., R.M., L.L.R., M.R., C.-Y.T. and T.S. contributed to data analysis and interpretation. A.B.-N. and O.M. contributed to data interpretation relevant to processes in ices. E.F.v.D. and C.W. contributed to data interpretation and writing of sections concerning interstellar oxygen. H.B., J.-J.B., P.B., J.D.K., B.F., S.A.F., A.K., U.M., B.M., T.O., H.R., J.H.W. and P.W. contributed to experiment design, calibration and data interpretation. All authors discussed the results, and commented on and revised the manuscript.

Author Information All ROSINA-DFMS data will be released to the PSA archive of ESA and to the PDS archive of NASA. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.B. (abieler@umich.edu).

METHODS

Data reduction. The integration time for all evaluated spectra is 20 s. A single mass spectrum consists of the signals of 512 individual pixels, which are arranged along the dispersive axis of the mass spectrometer. We take the gain degradation of each individual pixel over time into account by determining its gain with a calibration sequence dedicated to this task every few weeks.

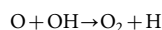
For every recorded mass spectrum, a third-degree polynomial is fitted to the baseline of the mass spectrum and subsequently subtracted from the signal. To avoid contamination in the baseline fit from the measured peaks, the centre of the mass spectrum (with the peaks) is ignored for those fitting purposes. Peaks are fitted with a Gaussian-shape curve. This introduces a systematic underestimation of the absolute magnitude of the signal, as at the 1% level and below the peaks are broader than Gaussian, but cancels out for the reported ratios of O_2/H_2O as both species are underestimated similarly. In a post-processing procedure, outliers are removed. An outlier occurs if one of the following criteria is met: (a) the signal amplitude changes by more than two orders of magnitudes between two measurements; (b) the centre of the fitted peak is not within a given window around the predicted position; (c) the width of the fitted peak curve is wider than a set limit.

For the CO measurements, the fraction contributed to the signal by CO fragments from CO_2 is subtracted. As DFMS measures one mass line at a time in high resolution mode, there are no simultaneous measurements of two species that are separated by more than 1 Da/e. To calculate ratios, the two measurements have to occur within 20 min; if no pair of measurements can be found within 20 min, the data point is ignored. In our error analysis we include uncertainties for the individual gain levels of the detector, the species branching ratios and their relative sensitivities as systematic errors. Individual pixel gains are treated as statistical error, as the peak position on the detector is not constant. A careful analysis confirmed that the O_2/H_2O ratio is independent of the peak centre positions.

Sample size. No statistical methods were used to predetermine sample size.

Spacecraft outgassing background. To clearly identify the measured O_2 as cometary in origin, all non-cometary sources of O_2 must be considered and excluded. The Rosetta spacecraft produces a neutral gas cloud of its own, mainly due to diffusion of volatiles out of spacecraft material and desorption of re-deposited volatiles from the spacecraft. For example, by changing the spacecraft attitude, different spacecraft elements are illuminated by the Sun, which then warm up and release condensed gas. The orange line in Fig. 1 shows the low-level signals from this spacecraft contamination for O_2 , S and CH_3OH , referred to as “background”. Those measurements were taken several days before the encounter with 67P. It is not possible to distinguish this background signal from any potential cometary signature with DFMS, but it has been well characterized before the arrival at 67P and is usually orders of magnitude lower than the measured O_2 signals²⁸. To keep the background influence as low as possible, we only considered mass spectra where both the O_2 and H_2O abundances are at least 5 times larger than the corresponding spacecraft contamination. Another potential source of O_2 is the oxidizer, N_2O_4 , used by the Rosetta spacecraft during thruster firings. Measurements taken shortly after a large thruster firing manoeuvre from June 2014 (still before arrival at the comet) show minor contaminations around 32 Da/e, but not directly affecting the O_2 peak (see green curve in Fig. 1). Although contamination from thruster firings is small, DFMS measurements are usually performed hours after thruster firings, in order to minimize influence thereof. Finally we can exclude the production of O_2 inside the instrument through a careful review of all oxygen-bearing molecules up to 150 Da/e, which could potentially fragment into O_2 in the DFMS electron impact ion source. Many minor species contain O_2 but these are too low in abundance to account for the large amount of O_2 detected. The remaining possibility is CO_2 , which is very abundant in the coma of 67P (ref. 6). However, owing to its molecular structure it only fragments into CO and O, not into O_2 (ref. 29). Finally, we exclude the production of O_2 from H_2O in the instrument. For 81 mass spectra taken from May to the end of June 2014 (before the encounter with 67P) we determine an O_2 abundance of $(0.18 \pm 0.07)\%$ relative to H_2O , which is a factor of 20 lower than the cometary values.

Coma chemistry. The production of O_2 from atomic oxygen in the neutral gas coma of 67P is possible through the reaction



Owing to the tenuous nature of the coma, only very few collisions are expected within the first 100 km above the nucleus surface³⁰. With increasing activity and therefore denser coma one would expect a higher O_2/H_2O ratio as both reactants (atomic oxygen and OH) are related to H_2O as well as the collision frequency. This is clearly not observed. Production of a significant amount of O_2 from coma chemistry hence seems unfeasible.

Correlation with H_2O . The measured O_2 signal shows a very strong dependence on radial distance (r) from the comet. It increases by roughly one order of magnitude when the radial distance from the comet decreases from 30 km to 10 km. This is in agreement with a predicted $1/r^2$ dependence of the number density profile of a non-reactive species. Examining the data further, we observe a strong correlation between H_2O and O_2 (see Fig. 2, Pearson correlation coefficient $R = 0.88$) for data from September 2014 to March 2015. This correlation indicates that O_2 and H_2O are both of a similar cometary origin. In contrast, there is no correlation between O_2 and H_2O for measurements taken before the arrival at the comet ($R = -0.01$). The observed temporal variations in the O_2/H_2O ratio are largely due to a nonlinear correlation between H_2O and O_2 for high water densities, where the O_2 ratio drops with increasing H_2O abundance. The correlation similarly supports the ruling out of CO_2 as a source of the O_2 , as previous studies^{6,23} and Fig. 2 show a lack of correlation between the abundance of H_2O and other species like CO, CO_2 and N_2 .

Radiolysis. The production of O_2 from water ice by radiolysis is the result of several reactions, where initially H, O and OH are produced, followed by subsequent rearrangement to form H_2 , HO_2 and H_2O_2 and ultimately O_2 (ref. 2). In a systematic study of the irradiation of pure cubic crystalline water ice at 10 K, it was estimated that comets in the Oort cloud would reach an equilibrium concentration of about 0.6% of H_2O_2 after $\sim 10^5$ years, and that the concentration of O_2 would be significantly lower³¹. As stated by the authors, the relative proportion of O_2 and H_2O_2 in actual astrophysical ices can be different as the presence of H_2O_2 ices may enhance the production of O_2 and vice versa.

Cosmic rays consist of energetic particles (mostly H^+) that can penetrate inside the cometary nucleus. Their penetration depth depends on their energy. The more energetic a cosmic ray, the deeper it deposits its energy into the cometary nucleus. Only the most energetic components, the Galactic cosmic rays (GCRs), can penetrate deeply into the cometary body. Today, the GCR flux peaks at energies around a few hundreds of MeV. At energies higher than ~ 1 GeV, the GCR flux drops exponentially, the flux at 10 GeV being approximately 100 times lower than the flux at 1 GeV. GCRs with energies near the GCR maximum of flux typically deposit their energy in the nucleus at depths of the order of metres to tens of metres³². There is some uncertainty on the evolution of the cosmic ray flux over the history of the Solar System related to the occurrence of supernovae which can drastically increase the cosmic ray flux at high energies (up to \sim PeV) during relatively short time periods (of the order of 10^5 years) or to the orbital motion of the Sun and of the Galaxy. There are also uncertainties on the cross-sections, the relative proportion of high-Z cosmic rays, and the role of porosity and defects in the cometary ice³³.

During 67P's stay in the Kuiper belt (~ 4.5 billion years) GCRs may have produced a significant amount of O_2 . This O_2 is produced where the GCR energy is deposited, that is, in the first tens of metres below the comet surface. Once cometary objects enter the inner Solar System they lose (owing to their activity and depending on their perihelion distance) material from the surface in the range of several metres per orbit around the Sun. Therefore, as most products built up during the stay in the Kuiper belt reside in the outer few metres, O_2 produced by radiolysis should be released quickly on the first solar passages. 67P's perihelion distance has been within the orbit of Jupiter for the past 250 years, possibly even for more than 5,000 years, and since a close encounter with Jupiter in 1959, the perihelion distance of 67P has been about 1.3 au with an orbital period of 6.4 years (ref. 34). Accumulated over the last perihelion passages we can assume that 67P has lost hundreds of metres from its surface.

However, O_2 molecules may have diffused into the comet and penetrated into deeper regions of the cometary nucleus. According to the estimated erosion of the comet, this requires significant diffusion of the O_2 molecules. Such efficient diffusion would dilute the produced O_2 , resulting in a lower O_2/H_2O ratio hardly compatible with the elevated amount of O_2 observed by DFMS.

Recent radiolysis. The flux of the low energy component such as solar energetic particles (SEP) is several orders of magnitude higher than the flux of GCRs. The SEP flux peaks at energies around a few tens of MeV and only penetrates the first centimetres of the comet³². Considering the comet erosion rate, the production of O_2 by low energy cosmic rays would have to be very fast to account for the observed amount of O_2 , requiring unrealistically high O_2 production rates. Consequently, it appears unlikely that most of the O_2 in 67P has been built up over the past few years.

O_2 production through surface sputtering. It has been shown that sputtering of refractory materials from the cometary surface due to the solar wind is occurring at 67P (ref. 35) and a clear difference between the southern and northern latitudes in the measured abundances of the sputtered species was demonstrated. This apparent spatial difference is explained via the asymmetry of the neutral coma, with higher number densities in the northern hemisphere that is preferentially exposed to the Sun for the time span under study. The solar wind, which is responsible for

the sputtering, is therefore attenuated more efficiently by these denser parts of the coma and thus has limited or no access to the surface. We find that the O_2/H_2O ratio is independent of latitude and relatively constant over a period of 7 months. Furthermore, the major part of the top surface of 67P accessible by the solar wind does not contain any water ice⁹. This suggests that sputtering cannot be the main source of the detected molecular oxygen. Moreover, the sputter yields by solar wind ions are orders of magnitude too low to explain the observed amount of O_2 .

Code availability. The data reduction software was written in the Julia and Python programming languages and is available upon request from A.B.

28. Schläppi, B. *et al.* Influence of spacecraft outgassing on the exploration of tenuous atmospheres with in situ mass spectrometry. *J. Geophys. Res.* **115**, A12313 (2010).
29. Tian, C. & Vidal, C. R. Electron impact dissociative ionization of CO_2 : measurements with a focusing time-of-flight mass spectrometer. *J. Chem. Phys.* **108**, 927 (1998).
30. Fuselier, S. A. *et al.* ROSINA/DFMS and IES observations of 67P: ion-neutral chemistry in the coma of a weakly outgassing comet. *Astron. Astrophys.* <http://dx.doi.org/10.1051/0004-6361/201526210> (2015).
31. Zheng, W., Jewitt, D. & Kaiser, R. I. Formation of hydrogen, oxygen and hydrogen peroxide in electron-irradiated crystalline water ice. *Astrophys. J.* **639**, 534–548 (2006).
32. Cooper, J. F., Christian, E. R. & Johnson, R. E. Heliospheric cosmic ray irradiation of Kuiper belt comets. *Adv. Space Res.* **21**, 1611–1614 (1998).
33. Grieves, G. A. & Orlando, T. M. The importance of pores in the electron stimulated production of D_2 and O_2 in low temperature ice. *Surf. Sci.* **593**, 180–186 (2005).
34. Maquet, L. The recent dynamical history of comet 67P/Churyumov-Gerasimenko. *Astron. Astrophys.* **579**, A78 (2015).
35. Wurz, P. *et al.* Solar wind sputtering of dust on the surface of 67P/Churyumov-Gerasimenko. *Astron. Astrophys.* <http://dx.doi.org/10.1051/0004-6361/201525980> (2015).

Loophole-free Bell inequality violation using electron spins separated by 1.3 kilometres

B. Hensen^{1,2}, H. Bernien^{1,2†}, A. E. Dréau^{1,2}, A. Reiserer^{1,2}, N. Kalb^{1,2}, M. S. Blok^{1,2}, J. Ruitenbergh^{1,2}, R. F. L. Vermeulen^{1,2}, R. N. Schouten^{1,2}, C. Abellán³, W. Amaya³, V. Pruneri^{3,4}, M. W. Mitchell^{3,4}, M. Markham⁵, D. J. Twitchen⁵, D. Elkouss¹, S. Wehner¹, T. H. Tamini^{1,2} & R. Hanson^{1,2}

More than 50 years ago¹, John Bell proved that no theory of nature that obeys locality and realism² can reproduce all the predictions of quantum theory: in any local-realist theory, the correlations between outcomes of measurements on distant particles satisfy an inequality that can be violated if the particles are entangled. Numerous Bell inequality tests have been reported^{3–13}; however, all experiments reported so far required additional assumptions to obtain a contradiction with local realism, resulting in ‘loopholes’^{13–16}. Here we report a Bell experiment that is free of any such additional assumption and thus directly tests the principles underlying Bell’s inequality. We use an event-ready scheme^{17–19} that enables the generation of robust entanglement between distant electron spins (estimated state fidelity of 0.92 ± 0.03). Efficient spin read-out avoids the fair-sampling assumption (detection loophole^{14,15}), while the use of fast random-basis selection and spin read-out combined with a spatial separation of 1.3 kilometres ensure the required locality conditions¹³. We performed 245 trials that tested the CHSH–Bell inequality²⁰ $S \leq 2$ and found $S = 2.42 \pm 0.20$ (where S quantifies the correlation between measurement outcomes). A null-hypothesis test yields a probability of at most $P = 0.039$ that a local-realist model for space-like separated sites could produce data with a violation at least as large as we observe, even when allowing for memory^{16,21} in the devices. Our data hence imply statistically significant rejection of the local-realist null hypothesis. This conclusion may be further consolidated in future experiments; for instance, reaching a value of $P = 0.001$ would require approximately 700 trials for an observed $S = 2.4$. With improvements, our experiment could be used for testing less-conventional theories, and for implementing device-independent quantum-secure communication²² and randomness certification^{23,24}.

We consider a Bell test in the form proposed by Clauser, Horne, Shimony and Holt (CHSH)²⁰ (Fig. 1a). The test involves two boxes labelled A and B. Each box accepts a binary input (0 or 1) and subsequently delivers a binary output (+1 or –1). In each trial of the Bell test, a random input bit is generated on each side and input to the respective box. The random input bit triggers the box to produce an output value that is recorded. The test concerns correlations between the output values (labelled x and y for boxes A and B, respectively) and the input bits (labelled a and b for A and B, respectively) generated within the same trial.

The discovery made by Bell is that in any theory of physics that is both local (physical influences do not propagate faster than light) and realistic (physical properties are defined before, and independent of, observation) these correlations are bounded more strongly than they are in quantum theory. In particular, if the input bits can be considered free random variables (condition of ‘free will’) and the boxes are

sufficiently separated such that locality prevents communication between the boxes during a trial, then the following inequality holds under local realism:

$$S = |\langle x \cdot y \rangle_{(0,0)} + \langle x \cdot y \rangle_{(0,1)} + \langle x \cdot y \rangle_{(1,0)} - \langle x \cdot y \rangle_{(1,1)}| \leq 2 \quad (1)$$

where $\langle x \cdot y \rangle_{(a,b)}$ denotes the expectation value of the product of x and y for input bits a and b . (A mathematical formulation of the concepts underlying Bell’s inequality is found in, for example, ref. 25.)

Quantum theory predicts that the Bell inequality can be significantly violated in the following setting. We add one particle, for example an electron, to each box. The spin degree of freedom of the electron forms a two-level system with eigenstates $|\uparrow\rangle$ and $|\downarrow\rangle$. For each trial, the two spins are prepared into the entangled state $|\psi^-\rangle = (|\uparrow\downarrow\rangle - |\downarrow\uparrow\rangle)/\sqrt{2}$. The spin in box A is then measured along direction Z (for input bit $a = 0$) or X (for $a = 1$) and the spin in box B is measured along $(-Z + X)/\sqrt{2}$ (for $b = 0$) or $(-Z - X)/\sqrt{2}$ (for $b = 1$). If the measurement outcomes are used as outputs of the boxes, then quantum theory predicts a value of $S = 2\sqrt{2}$, which shows that the combination of locality and realism is fundamentally incompatible with the predictions of quantum mechanics.

Bell’s inequality provides a powerful recipe for probing fundamental properties of nature: all local-realist theories that specify where and when the free random input bits and the output values are generated can be experimentally tested against it.

Violating Bell’s inequality with entangled particles poses two main challenges: excluding any possible communication between the boxes (locality loophole¹³) and guaranteeing efficient measurements (detection loophole^{14,15}). First, if communication is possible, a box can in principle respond using knowledge of both input settings, rendering the Bell inequality invalid. The locality conditions thus require boxes A and B and their respective free-input-bit generations to be separated in such a way that signals travelling at the speed of light (the maximum allowed under special relativity) cannot communicate the local input setting of box A to box B, before the output value of box B has been recorded, and vice versa. Second, disregarding trials in which a box does not produce an output bit (that is, assuming fair sampling) would allow the boxes to select trials on the basis of the input setting. The fair sampling assumption thus opens a detection loophole^{14,15}: the selected subset of trials may show a violation even though the set of all trials may not.

The locality loophole has been addressed with pairs of photons separated over a large enough distance, in combination with fast settings changes⁴ and later with settings determined by fast random number generators^{5,9}. However, these experiments left open the detection loophole, owing to imperfect detectors and inevitable photon loss during the spatial distribution of entanglement. The detection loophole has been closed in different experiments^{6–8,10–12}, but these did not

¹QuTech, Delft University of Technology, PO Box 5046, 2600 GA Delft, The Netherlands. ²Kavli Institute of Nanoscience Delft, Delft University of Technology, PO Box 5046, 2600 GA Delft, The Netherlands. ³ICFO-Institut de Ciències Fotoniques, The Barcelona Institute of Science and Technology, 08860 Castelldefels (Barcelona), Spain. ⁴ICREA-Institució Catalana de Recerca i Estudis Avançats, Lluís Companys 23, 08010 Barcelona, Spain. ⁵Element Six Innovation, Fermi Avenue, Harwell Oxford, Didcot, Oxfordshire OX11 0QR, UK. [†]Present address: Department of Physics, Harvard University, Cambridge, Massachusetts 02138, USA.

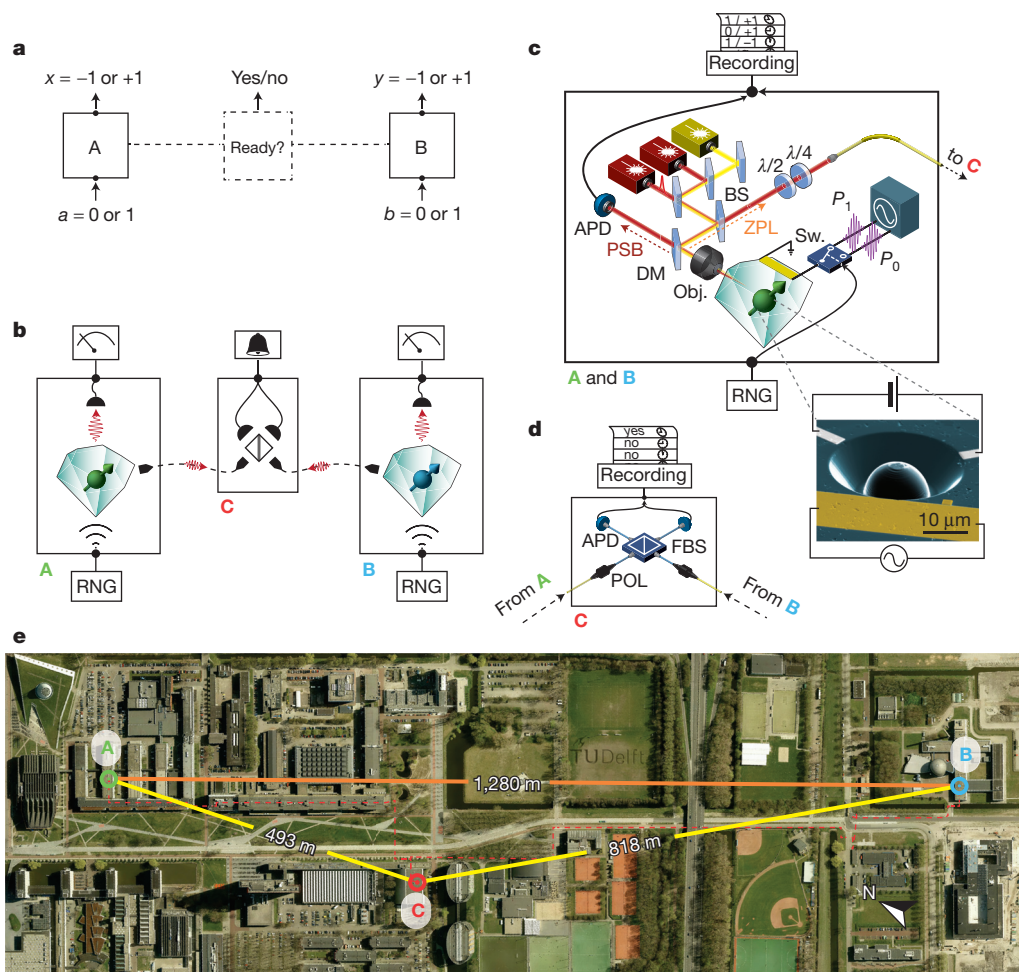


Figure 1 | Bell-test schematic and experimental realization. **a**, Bell-test set-up: two boxes, A and B, accept binary inputs (a, b) and produce binary outputs (x, y). In an event-ready scenario, an additional box C gives a binary output signalling that A and B were successfully prepared. **b**, Experimental realization. The set-up consists of three separate laboratories, A, B and C. The boxes at locations A and B each contain a single NV centre in diamond. A quantum random-number generator (RNG) is used to provide the input. The NV electronic spin is read out in a basis that depends on the input bit, and the resultant signal provides the output. A box at location C records the arrival of single photons that were previously emitted by, and entangled with, the spins at A and B. **c**, Experimental set-up at A and B. The NV centre is located in a low-temperature confocal microscope (Obj.). Depending on the output of the RNG, a fast switch (Sw.) transmits one of two different microwave pulses (P_0 and P_1) into a gold line deposited on the diamond surface (inset, scanning

electron microscope image). Pulsed red and yellow lasers are used to resonantly excite the optical transitions of the NV centre. The emission (dashed arrows) is spectrally separated into an off-resonant part (phonon side band, PSB) and a resonant part (zero-phonon line, ZPL), using a dichroic mirror (DM). The PSB emission is detected with a single-photon counter (APD). The ZPL emission is transmitted through a beam-splitter (BS, reflection $\leq 4\%$) and wave plates ($\lambda/2$ and $\lambda/4$), and sent to location C through a single-mode fibre. **d**, Set-up at location C. The fibres from A and B are connected to a fibre-based beam splitter (FBS) after passing a fibre-based polarizer (POL). Photons in the output ports are detected and recorded. **e**, Aerial photograph of the campus of Delft University of Technology indicating the distances between locations A, B and C. The red dotted line marks the path of the fibre connection. Aerial photograph by Slagboom en Peeters Luchtfotografie BV.

close the locality loophole. So far, no experiment has closed all the loopholes simultaneously.

A Bell test that closes all experimental loopholes at the same time—commonly referred to as a loophole-free Bell test^{15,19}—is of foundational importance to the understanding of nature. In addition, a loophole-free Bell test is a critical component for device-independent quantum security protocols²² and randomness certification^{23,24}. In such adversarial scenarios, all loopholes are ideally closed because they allow for security breaches in the system²⁶.

One approach for realizing a loophole-free set-up was proposed by Bell himself¹⁷. The key idea is to record an additional signal (dashed box in Fig. 1a) to indicate whether the required entangled state was successfully shared between A and B, that is, whether the boxes were ready to be used for a trial of the Bell test. By conditioning the validity of a Bell-test trial on this event-ready signal, failed entanglement distribution events are excluded upfront from being used in the Bell test.

We implemented an event-ready Bell set-up^{18,19} with boxes that use the electronic spin associated with a single nitrogen-vacancy (NV) defect centre in a diamond chip (Fig. 1b). The diamond chips are mounted in closed-cycle cryostats ($T = 4$ K) located in distant laboratories named A and B (Fig. 1c). We control the electronic spin state of each NV centre with microwave pulses applied to on-chip striplines (Fig. 1c, inset). The spins are initialized through optical pumping and read out along the Z axis via spin-dependent fluorescence²⁷. The read-out relies on resonant excitation of a spin-selective cycling transition (12-ns lifetime), which causes the NV centre to emit many photons when it is in the bright $m_s = 0$ spin state, while it remains dark when it is in either of the $m_s = \pm 1$ states. We assign the value +1 ($m_s = 0$) to the output if we record at least one photo-detector count during the read-out window, and the value -1 ($m_s = \pm 1$) otherwise. Read-out in a rotated basis is achieved by first rotating the spin, followed by read-out along Z.

We generate entanglement between the two distant spins by entanglement swapping¹⁸ in the Barrett–Kok scheme^{28,29} using a third loca-

tion C (roughly midway between A and B; see Fig. 1e). First we entangle each spin with the emission time of a single photon (time-bin encoding). The two photons are then sent to location C, where they are overlapped on a beam-splitter and subsequently detected. If the photons are indistinguishable in all degrees of freedom, then the observation of one early and one late photon in different output ports projects the spins at A and B into the maximally entangled state $|\psi^-\rangle = (|\uparrow\downarrow\rangle - |\downarrow\uparrow\rangle)/\sqrt{2}$, where $m_s = 0 \equiv |\uparrow\rangle$ and $m_s = -1 \equiv |\downarrow\rangle$. These detections herald the successful preparation and play the role of the event-ready signal in Bell's proposed set-up. As can be seen in the space-time diagram in Fig. 2a, we ensure that this event-ready signal is space-like separated from the random input-bit generation at locations A and B.

The separation of the spins by 1,280 m defines a 4.27- μ s time window during which the local events at A and B are space-like separated from each other (see the space-time diagram in Fig. 2b). To comply with the locality conditions of the Bell test, the choice of measurement bases and the measurement of the spins should be performed within this time window. For the basis choice we use fast random-number generators with real-time randomness extraction³⁰. We reserve 160 ns for the random basis choice, during which time one extremely random bit is generated from 32 partially random raw bits (Supplementary

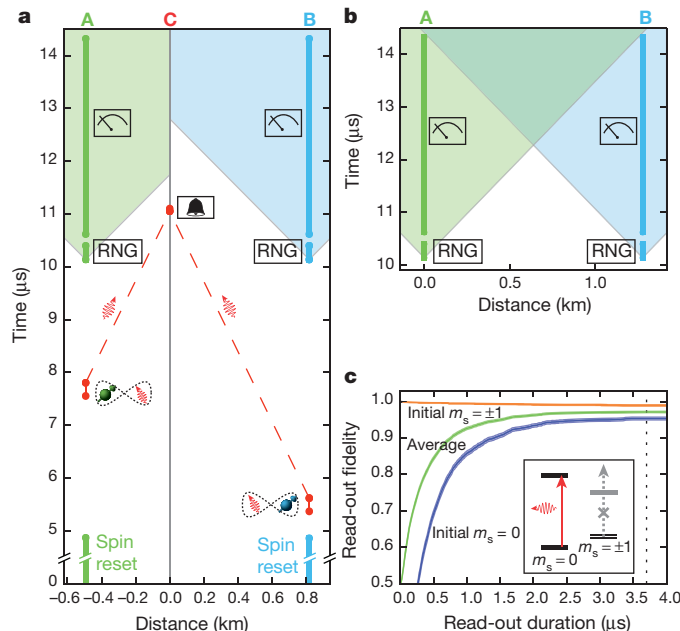


Figure 2 | Space-time analysis of the experiment. **a**, Space-time diagram of a single repetition of the entanglement generation. The x axis denotes the distance along the lines AC and CB. After spin initialization, spin-photon entanglement is generated, such that the two photons from A and B arrive simultaneously at C where the detection time of the photons is recorded. Successful preparation of the spins is signalled (bell symbol) by a specific coincidence detection pattern. Independent of the event-ready signal, the set-ups at locations A and B choose a random basis (RNG symbol), rotate the spin accordingly and start the optical spin read-out (measurement symbol). Vertical bars indicate durations. The event-ready signal lies outside the future light cone (coloured regions) of the random basis choices of A and B. **b**, Space-time diagram of the Bell test. The x axis denotes the distance along the line AB. The read-out on each side is completed before any light-speed signal can communicate the basis choice from the other side. The uncertainty in the depicted event times and locations is much smaller than the symbol size. **c**, Single-shot spin read-out fidelity at location A as a function of read-out duration (set by the latest time that detection events are taken into account). Blue (orange) line, fidelity of outcome +1 (−1) when the spin is prepared in $m_s = 0$ ($m_s = \pm 1$); green line, average read-out fidelity; dotted line, read-out duration used (3.7 μ s). The inset shows the relevant ground and excited-state levels (not to scale).

Information). The random bit sets the state of a fast microwave switch that selects one out of two preprogrammed microwave pulses implementing the two possible read-out bases (Fig. 1c). Adding the durations of each of the above steps yields a maximum time from the start of the basis choice to the start of the read-out of 480 ns. We choose the read-out duration to be 3.7 μ s, which leaves 90 ns to cover any uncertainty in the distance between the laboratories and the synchronization of the set-up (estimated total error is at most 16 ns; see Supplementary Information). For this read-out duration, the combined initialization and single-shot read-out fidelity of sample A is $(97.1 \pm 0.2)\%$ (Fig. 2c); sample B achieves $(96.3 \pm 0.3)\%$. In summary, the use of the event-ready scheme enables us to comply with the strict locality conditions of the Bell set-up by using photons to distribute entanglement, while simultaneously using the single-shot nature of the spin read-out to close the detection loophole.

Before running the Bell test we first characterized the set-up and the preparation of the spin–spin entangled state. Figure 3a displays correlation measurements on the entangled spin-photon states to be used for the entanglement swapping. For both locations A and B we observe near-unity correlations between spin state and photon time bin when spin read-out errors are accounted for. We then estimate the degree of indistinguishability of the single photons emitted at locations A and B in a Hong–Ou–Mandel³¹ two-photon-interference experiment at location C, that is, after the photons have travelled through a combined length of 1.7 km of fibre. Using the observed two-photon interference contrast of 0.90 ± 0.06 and the spin-photon correlation data, we estimate that the fidelity to the ideal state $|\psi^-\rangle$ of the spin–spin entangled states generated in our set-up is 0.92 ± 0.03 (Supplementary Information). Combined with measured read-out fidelities, the generated entangled state is thus expected to violate the CHSH–Bell inequality with $S = 2.30 \pm 0.07$.

As a final characterization we ran the full Bell sequence including random number generation and fast read-out, but with co-linear measurement bases (ZZ and XX) such that spin–spin correlations could be observed with optimal contrast. To test the fast basis selection and rotation, the Z (X) basis measurements are randomly performed along the +Z (+X) and −Z (−X) axis. The observed correlations, shown in Fig. 3c (orange bars), are consistent with the estimated quantum state and the independently measured read-out fidelities (dotted bars), which confirms that the set-up is performing as expected and that the desired entangled state is generated.

We find a success probability per entanglement generation attempt of about 6.4×10^{-9} , which yields slightly more than one event-ready signal per hour. Compared to our previous heralded entanglement experiments over 3 m (ref. 29), this probability is reduced, mainly owing to additional photon loss (8 dB km^{−1}) in the 1.7-km optical fibre. To ensure the required long-term operation, we exploit active stabilization on different relevant timescales via automated feedback loops (Supplementary Information). We note that the distance between the entangled electrons is nearly two orders of magnitude larger than it was in any previous experiment^{7,10,29,32} with entangled matter systems.

Using the results of the characterization measurements we determine the optimal read-out bases for our Bell test. A numerical optimization yields the following angles for the read-out bases with respect to Z: 0 (for $a = 0$), $+\pi/2$ (for $a = 1$), $-3\pi/4 - \varepsilon$ (for $b = 0$) and $3\pi/4 + \varepsilon$ (for $b = 1$), with $\varepsilon = 0.026\pi$. Adding the small angle ε is beneficial because of the stronger correlations in ZZ compared to XX. Furthermore, we use the characterization data to determine the time window for valid photon-detection events at location C to optimally reject reflected laser light and detector dark counts. We choose this window conservatively to optimize the entangled-state fidelity at the cost of a reduced data rate. These settings are then fixed and used throughout the actual Bell test. As a final optimization we replaced the photo-detectors at location C with the best set we had available.

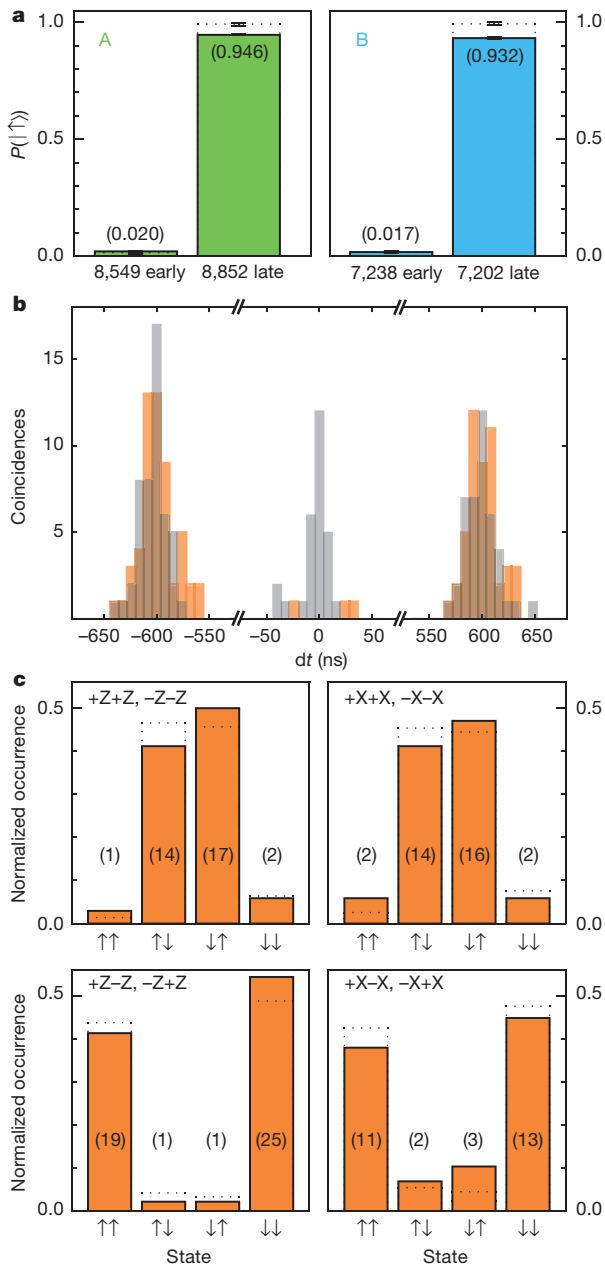


Figure 3 | Characterization of the set-up and the entangled state. **a**, The probability to obtain spin state $|\uparrow\rangle$ at location A (left panel) or B (right panel) when a single photon is detected in the early or late time bin at location C. In the left (right) panel, only emission from A (B) was recorded. Dotted bars are corrected for finite spin read-out fidelity and yield remaining errors of $1.4\% \pm 0.2\%$ ($1.6\% \pm 0.2\%$) and $0.8\% \pm 0.4\%$ ($0.7\% \pm 0.4\%$) for early and late detection events, respectively, from set-up A (B). These errors include imperfect rejection of the excitation laser pulses, detector dark counts, microwave-pulse errors and off-resonant excitation of the NV. **b**, Two-photon quantum interference signal, with dt the time between the two photo-detection events. When the NV centres at A and B emit indistinguishable photons, coincident detections of two photons, one in each output arm of the beam-splitter at C, are expected to vanish. The observed contrast between the cases of indistinguishable (orange) and distinguishable (grey) photons (3 versus 28 events in the central peak) yields a visibility of $90 \pm 6\%$ (Supplementary Information). **c**, Characterization of the Bell set-up using (anti-)parallel read-out angles. The spins at A (left arrows on the x axis) and B (right arrows on the x axis) are read out along the $\pm Z$ axis (left panels) or the $\pm X$ axis (right panels). The numbers in brackets are the raw number of events. The dotted lines represent the expected correlations on the basis of the characterization measurements presented in **a** and **b** (Supplementary Information). The data yield a strict lower bound²⁹ on the state fidelity to $|\psi^-\rangle$ of 0.83 ± 0.05 . Error bars are 1 s.d.

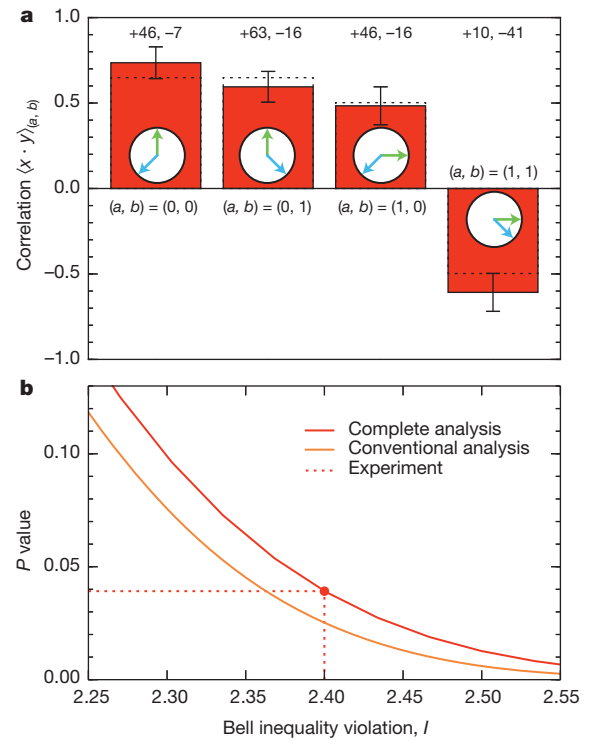


Figure 4 | Loophole-free Bell inequality violation. **a**, Summary of the data and the CHSH correlations. The read-out bases corresponding to the input values are indicated by the green (for A) and blue (for B) arrows. Dotted lines indicate the expected correlation on the basis of the spin read-out fidelities and the characterization measurements presented in Fig. 3 (Supplementary Information). Numbers above the bars represent the number of correlated and anti-correlated outcomes, respectively. Error bars shown are $\sqrt{(1 - \langle x \cdot y \rangle_{(a,b)}^2) / n_{(a,b)}}$, with $n_{(a,b)}$ the number of events with inputs (a, b) . **b**, Statistical analysis for $n = 245$ trials. For the null-hypothesis test performed (Supplementary Information), the dependence of the P value on the I value is shown (complete analysis, red). Here $I = 8(\frac{k}{n} - \frac{1}{2})$, with k the number of times $(-1)^{(a \cdot b)} x \cdot y = 1$. (For equal $n_{(a,b)}$, $I = S$ with S defined in equation (1).) A small P value indicates strong evidence against the null hypothesis. We find $k = 196$, which results in a rejection of the null hypothesis with a $P \leq 0.039$. For comparison, we also plot the P value for an analysis (conventional analysis, orange) assuming independent and identically distributed (i.i.d.) trials, Gaussian statistics, no memory and perfect random-number generators.

We ran 245 trials of the Bell test during a total measurement time of 220 h over a period of 18 days. Figure 4a summarizes the observed data, from which we find $S = 2.42$, in violation of the CHSH-Bell inequality $S \leq 2$. We quantify the significance of this violation for two different scenarios (see Fig. 4b). First, similar to previous work⁴⁻⁹, we analyse the data under the assumptions that the Bell trials are independent of each other, that the recorded random input bits have zero predictability and that the outcomes follow a Gaussian distribution. This analysis (which we term ‘conventional’) yields a standard deviation of 0.20 on S . In this case, the null hypothesis that a local-realist model for space-like separated sites describes our experiment is rejected with a P value of 0.019 (see Supplementary Information).

The assumptions made in the conventional analysis are not justified in a typical Bell experiment. For instance, although the locality conditions outlined earlier are designed to ensure independent operation during a single trial, the boxes can in principle have access to the entire history including results from all previous trials and adjust their output to it^{16,21}. Our second analysis (which we term ‘complete’) allows for arbitrary memory, takes the partial predictability of the random input bits into account and also makes no assumption about the probability distributions underlying the data (see Supplementary Information). In

this case, the null hypothesis that an arbitrary local-realist model of space-like separated sites governs our experiment is rejected with a P value of 0.039 (Fig. 4b). This P value might be further tightened in future experiments.

Our experiment realizes the first Bell test that simultaneously addresses both the detection loophole and the locality loophole. Being free of the experimental loopholes, the set-up tests local-realist theories of nature without introducing extra assumptions such as fair sampling, a limit on (sub-)luminal communication or the absence of memory in the set-up. Our observation of a statistically significant loophole-free Bell inequality violation thus indicates rejection of all local-realist theories that accept that the number generators produce a free random bit in a timely manner and that the outputs are final once recorded in the electronics.

Strictly speaking, no Bell experiment can exclude all conceivable local-realist theories, because it is fundamentally impossible to prove when and where free random input bits and output values came into existence¹³. Even so, our loophole-free Bell test opens the possibility to progressively bound such less-conventional theories: by increasing the distance between A and B (for example, to test theories with increased speed of physical influence); by using different random input bit generators (to test theories with specific free-will agents, for example, humans); or by repositioning the random input bit generators (to test theories where the inputs are already determined earlier, sometimes referred to as ‘freedom-of-choice’⁹). In fact, our experiment already enables tests of all models that predict that the random inputs are determined a maximum of 690 ns before we record them (Supplementary Information).

Combining the presented event-ready scheme with higher entangling rates (for example, through the use of optical cavities) provides prospects for the implementation of device-independent quantum key distribution²² and randomness certification^{23,24}. In combination with quantum repeaters, this might enable the realization of large-scale quantum networks that are secured through the very same counter-intuitive concepts that inspired one of the most fundamental scientific debates for 80 years^{1,2,25}.

Received 19 August; accepted 28 September 2015.

Published online 21 October 2015.

1. Bell, J. S. On the Einstein–Podolsky–Rosen paradox. *Physics* **1**, 195–200 (1964).
2. Einstein, A., Podolsky, B. & Rosen, N. Can quantum-mechanical description of physical reality be considered complete? *Phys. Rev.* **47**, 777–780 (1935).
3. Freedman, S. J. & Clauser, J. F. Experimental test of local hidden-variable theories. *Phys. Rev. Lett.* **28**, 938–941 (1972).
4. Aspect, A., Dalibard, J. & Roger, G. Experimental test of Bell’s inequalities using time-varying analyzers. *Phys. Rev. Lett.* **49**, 1804–1807 (1982).
5. Weihs, G., Jennewein, T., Simon, C., Weinfurter, H. & Zeilinger, A. Violation of Bell’s inequality under strict Einstein locality conditions. *Phys. Rev. Lett.* **81**, 5039–5043 (1998).
6. Rowe, M. A. *et al.* Experimental violation of a Bell’s inequality with efficient detection. *Nature* **409**, 791–794 (2001).
7. Matsukevich, D. N., Maunz, P., Moehring, D. L., Olmschenk, S. & Monroe, C. Bell inequality violation with two remote atomic qubits. *Phys. Rev. Lett.* **100**, 150404 (2008).
8. Ansmann, M. *et al.* Violation of Bell’s inequality in Josephson phase qubits. *Nature* **461**, 504–506 (2009).
9. Scheidl, T. *et al.* Violation of local realism with freedom of choice. *Proc. Natl Acad. Sci. USA* **107**, 19708–19713 (2010).
10. Hofmann, J. *et al.* Heralded entanglement between widely separated atoms. *Science* **337**, 72–75 (2012).
11. Giustina, M. *et al.* Bell violation using entangled photons without the fair-sampling assumption. *Nature* **497**, 227–230 (2013).

12. Christensen, B. G. *et al.* Detection-loophole-free test of quantum nonlocality, and applications. *Phys. Rev. Lett.* **111**, 130406 (2013).
13. Brunner, N., Cavalcanti, D., Pironio, S., Scarani, V. & Wehner, S. Bell nonlocality. *Rev. Mod. Phys.* **86**, 419–478 (2014).
14. Garg, A. & Mermin, N. D. Detector inefficiencies in the Einstein–Podolsky–Rosen experiment. *Phys. Rev. D* **35**, 3831–3835 (1987).
15. Eberhard, P. H. Background level and counter efficiencies required for a loophole-free Einstein–Podolsky–Rosen experiment. *Phys. Rev. A* **47**, R747–R750 (1993).
16. Barrett, J., Collins, D., Hardy, L., Kent, A. & Popescu, S. Quantum nonlocality, Bell inequalities, and the memory loophole. *Phys. Rev. A* **66**, 042111 (2002).
17. Bell, J. S. Atomic-cascade photons and quantum-mechanical nonlocality. *Comments Atom. Mol. Phys.* **9**, 121–126 (1980).
18. Żukowski, M., Zeilinger, A., Horne, M. A. & Ekert, A. K. “Event-ready-detectors” Bell experiment via entanglement swapping. *Phys. Rev. Lett.* **71**, 4287–4290 (1993).
19. Simon, C. & Irvine, W. T. M. Robust long-distance entanglement and a loophole-free Bell test with ions and photons. *Phys. Rev. Lett.* **91**, 110405 (2003).
20. Clauser, J. F., Horne, M. A., Shimony, A. & Holt, R. A. Proposed experiment to test local hidden-variable theories. *Phys. Rev. Lett.* **23**, 880–884 (1969).
21. Gill, R. D. Time, finite statistics, and Bell’s fifth position. In *Proc. Foundations of Probability and Physics 2* 179–206 (Vaxjö Univ. Press, 2003).
22. Acín, A. *et al.* Device-independent security of quantum cryptography against collective attacks. *Phys. Rev. Lett.* **98**, 230501 (2007).
23. Colbeck, R. *Quantum and Relativistic Protocols for Secure Multi-Party Computation*. PhD thesis, Univ. Cambridge (2007); <http://arxiv.org/abs/0911.3814>.
24. Pironio, S. *et al.* Random numbers certified by Bell’s theorem. *Nature* **464**, 1021–1024 (2010).
25. Bell, J. S. *Speakable and Unspeakable in Quantum Mechanics: Collected Papers on Quantum Philosophy* 2nd edn (Cambridge Univ. Press, 2004).
26. Gerhardt, I. *et al.* Experimentally faking the violation of Bell’s inequalities. *Phys. Rev. Lett.* **107**, 170404 (2011).
27. Robledo, L. *et al.* High-fidelity projective read-out of a solid-state spin quantum register. *Nature* **477**, 574–578 (2011).
28. Barrett, S. D. & Kok, P. Efficient high-fidelity quantum computation using matter qubits and linear optics. *Phys. Rev. A* **71**, 060310 (2005).
29. Bernien, H. *et al.* Heralded entanglement between solid-state qubits separated by three metres. *Nature* **497**, 86–90 (2013).
30. Abellan, C., Amaya, W., Mitrani, D., Pruneri, V. & Mitchell, M. W. Generation of fresh and pure random numbers for loophole-free Bell tests. Preprint available at <http://arxiv.org/abs/1506.02712>.
31. Hong, C. K., Ou, Z. Y. & Mandel, L. Measurement of subpicosecond time intervals between two photons by interference. *Phys. Rev. Lett.* **59**, 2044–2046 (1987).
32. Ritter, S. *et al.* An elementary quantum network of single atoms in optical cavities. *Nature* **484**, 195–200 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank A. Acín, A. Aspect, P. Bierhorst, A. Doherty, R. Gill, P. Grünwald, M. Giustina, L. Mancinska, J. E. Mooij, T. Vidick, H. Weinfurter and Y. Zhang for discussions and/or reading our manuscript, and M. Blauw, P. Dorenbos, R. de Stefano, C. Tiberius, T. Versluis, R. Zwagerman and Facilitair Management and Vastgoed for help with the realization of the laboratories and the optical fibre connections. We acknowledge support from the Dutch Organization for Fundamental Research on Matter (FOM), the Dutch Technology Foundation (STW), the Netherlands Organization for Scientific Research (NWO) through a VENI grant (T.H.T.) and a VIDI grant (S.W.), the Defense Advanced Research Projects Agency QuASAR program, the Spanish MINECO project MAGO (reference FIS2011-23520) and Explora Ciencia (reference FIS2014-62181-EXP), the European Regional Development Fund (FEDER) grant TEC2013-46168-R, Fundacio Privada CELLEX, FET Proactive project QUIC and the European Research Council through projects AQUMET and HYSOCORE.

Author Contributions B.H., H.B. and R.H. devised the experiment. B.H., H.B., A.E.D., A.R., M.S.B., J.R., R.F.L.V. and R.N.S. built and characterized the experimental set-up. M.W.M., C.A. and V.P. designed the quantum random-number generators (QRNGs), M.W.M. and C.A. designed the randomness extractors, and W.A. and C.A. built the interface electronics and the QRNG optics, the latter with advice from V.P. C.A. and M.W.M. designed and implemented the QRNG statistical metrology. C.A. designed and implemented the QRNG output tests. M.M. and D.J.T. grew and prepared the diamond device substrates. H.B. and M.S.B. fabricated the devices. B.H., H.B., A.E.D., A.R. and N.K. collected and analysed the data, with support from T.H.T. and R.H. D.E. and S.W. performed the theoretical analysis. B.H., A.R., T.H.T., D.E., S.W. and R.H. wrote the manuscript. R.H. supervised the project.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to R.H. (r.hanson@tudelft.nl).

Organometallic palladium reagents for cysteine bioconjugation

Ekaterina V. Vinogradova^{1*}, Chi Zhang^{1*}, Alexander M. Spokoyny^{1†}, Bradley L. Pentelute¹ & Stephen L. Buchwald¹

Reactions based on transition metals have found wide use in organic synthesis, in particular for the functionalization of small molecules^{1,2}. However, there are very few reports of using transition-metal-based reactions to modify complex biomolecules^{3,4}, which is due to the need for stringent reaction conditions (for example, aqueous media, low temperature and mild pH) and the existence of multiple reactive functional groups found in biomolecules. Here we report that palladium(II) complexes can be used for efficient and highly selective cysteine conjugation (bioconjugation) reactions that are rapid and robust under a range of bio-compatible reaction conditions. The straightforward synthesis of the palladium reagents from diverse and easily accessible aryl halide and trifluoromethanesulfonate precursors makes the method highly practical, providing access to a large structural space for protein modification. The resulting aryl bioconjugates are stable towards acids, bases, oxidants and external thiol nucleophiles. The broad utility of the bioconjugation platform was further corroborated by the synthesis of new classes of stapled peptides and antibody–drug conjugates. These palladium complexes show potential as benchtop reagents for diverse bioconjugation applications.

Post-translational modifications greatly expand the function of proteins⁵. Chemists aim to mimic the success of such natural transformations through the development of chemo- and regioselective reactions of proteins. The diversity of potentially reactive functional groups present in biomolecules (for example, amides, acids, alcohols, amines) combined with the requirement for fast kinetics and mild reaction conditions (for example, aqueous solvent, pH 6–8, temperature $T < 37^\circ\text{C}$) make challenging the development of new techniques to functionalize proteins. Nevertheless, methods have emerged for bioconjugation with natural and unnatural amino acids in protein molecules^{6,7}. Cysteine is a key residue for the chemical modification of proteins owing to the unique reactivity of the thiol functional group and the low abundance of cysteine residues in naturally occurring proteins^{8,9}. Michael addition to maleimides and $\text{S}_{\text{N}}2$ reaction with alkyl halides are commonly used for cysteine modification. The resulting conjugates tend to decompose in the presence of external bases or thiol nucleophiles¹⁰, which prompted the recent development of advanced cysteine bioconjugations for the improved stability of conjugates (see ref. 11, and references therein).

The ability to achieve high levels of chemo- and regioselectivity through the judicious choice of metal and ligand design suggest that metal-mediated processes could be very attractive for the development of new bioconjugations. Existing metal-based transformations often rely on the use of functional handles¹² or unnatural amino acids, such as 4-iodophenylalanine and aldehyde- or alkyne-containing amino acids^{3,4,13}, and require high concentrations (mM) of derivatizing agents, which can cause off-target reactivity or purification problems. We considered that palladium complexes resulting from the oxidative addition of aryl halides or trifluoromethanesulfonates¹⁴ could be used

for the transfer of aryl groups to cysteine residues in proteins (Fig. 1a). (For existing transition-metal-catalysed C–S bond-forming reactions, see ref. 15.) The efficiency and selectivity of the proposed reaction with the highly active palladium species may be hampered by the presence of a variety of functional groups within complex biomolecules. Furthermore, the presence of free thiols has been previously shown to inhibit palladium-catalysed cross-coupling reactions on peptides¹⁶, while Pd(II)-complexes have also been shown to exhibit protease-like behaviour with certain peptides¹⁷. However, we envisioned that the careful choice of ligand would provide stable, yet highly reactive reagents for the desired transformations (Fig. 1b), while the interaction between the soft nucleophile cysteine thiol and the aryl palladium(II) species would guide its selectivity.

We began our study with a palladium–tolyl complex (**1A-OTf**; Fig. 1c) using 2-dicyclohexylphosphino-2',6'-diisopropoxybiphenyl (RuPhos) as the ligand and trifluoromethanesulfonate as the counterion. A model peptide (**P1**; Fig. 1c) was used for the optimization of the reaction conditions and for the exploration of the substrate scope. Full conversion of the starting peptide to the corresponding aryl product was observed in less than 5 min at low micromolar concentrations of reagents (Fig. 1c). Further, the reaction was selective for cysteine. No reaction was observed using a control peptide with cysteine mutated to serine (Supplementary Information), in contrast to the palladium-mediated protein allylation, which is selective for tyrosine (O-allylation) over lysine and cysteine (N- and S-allylation). (The new aryl–palladium reagents are less electrophilic than the allylpalladium species used in ref. 18, which tunes their selectivity towards cysteine residues while making them completely unreactive towards alcohol-based species like tyrosine.) These results highlight the importance of choosing the right ligand, which facilitates C–S reductive elimination and, together with the overall electrophilicity of the palladium centre, tunes the selectivity of the transformation.

Most cysteine conjugation reactions operate at nearly neutral to slightly basic pH. Further evaluation of the reaction conditions using palladium reagents revealed quantitative conversion of the starting peptide to the corresponding S-aryl cysteine conjugate within a broad pH range (5.5–8.5) using common organic co-solvents (5% of *N,N*-dimethylformamide (DMF), dimethylsulfoxide (DMSO), acetonitrile (CH_3CN)) in various buffers (Supplementary Information, Supplementary Table 1). Remarkably, even in 0.1% trifluoroacetic acid (TFA) solution (pH 2.0) the reaction yielded 59% of the S-arylated product after 7 h. The process was also compatible with the protein disulfide reducing agent *tris*(2-carboxyethyl)phosphine (TCEP) that has been shown to hamper bioconjugations by reacting with maleimide and α -haloacyl groups¹⁹.

The palladium-mediated conjugation is fast, and complete product formation occurs within 15 s at 4°C . The reaction rate was estimated by competition experiments against the commonly used *N*-methyl maleimide cysteine ligation. (The reported²⁰ kinetics for maleimide conjugation are in the range 10^3 – $10^4 \text{ M}^{-1} \text{ s}^{-1}$ at pH 7.5.) At pH 7.5

¹Department of Chemistry, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [†]Present address: Department of Chemistry and Biochemistry, University of California, Los Angeles, California 90095, USA.

*These authors contributed equally to this work.

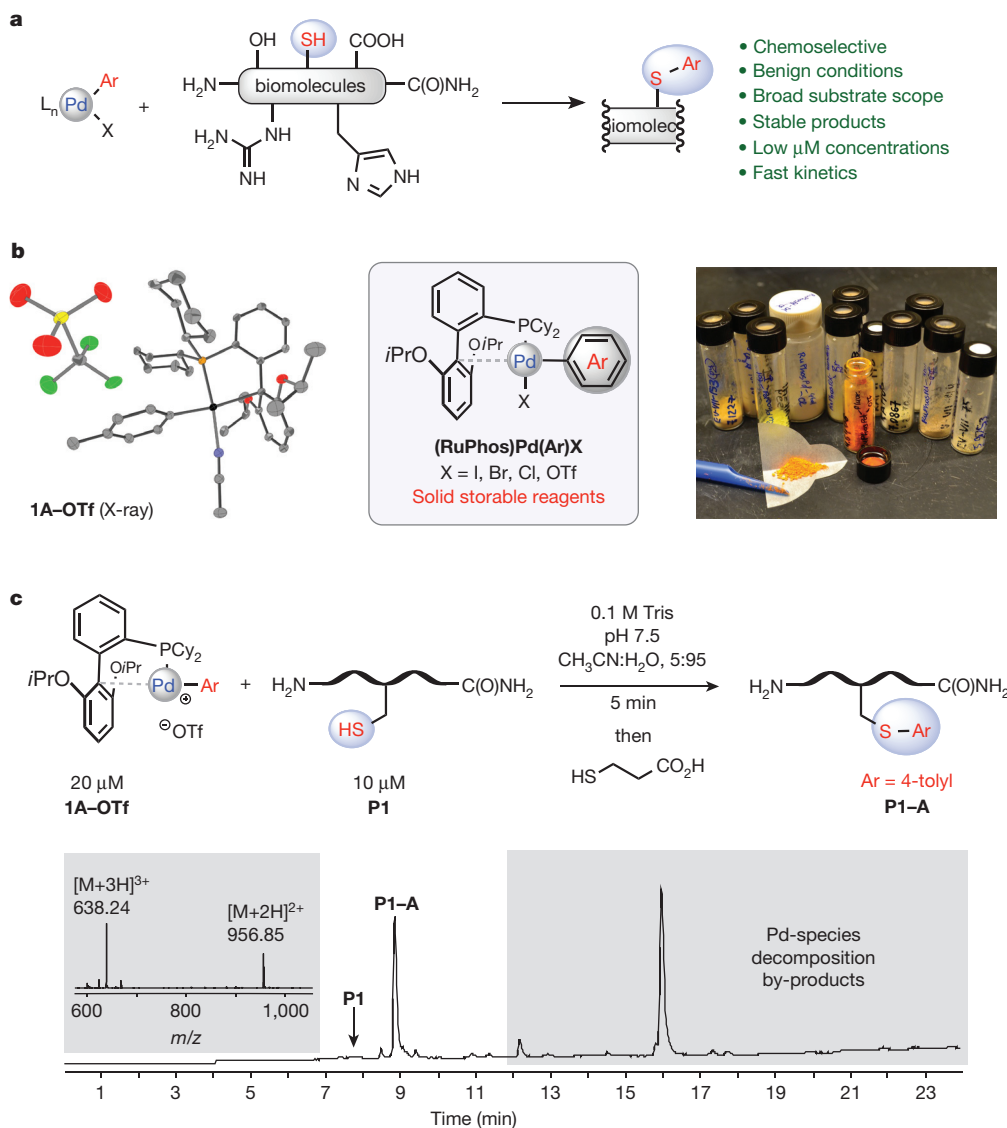


Figure 1 | Organometallic palladium reagents for cysteine modification.

Figure shows general strategy and model studies. **a**, Proposed cysteine bioconjugation using palladium reagents. **b**, X-ray structure, storage and operational simplicity of palladium reagents for cysteine bioconjugation. Left, crystals of **1A-OTf** for X-ray diffraction were obtained by vapour diffusion of a diethyl ether/acetonitrile solution of **1A-OTf** with pentane. Thermal ellipsoid plots are drawn at 50% probability, hydrogen atoms are omitted for clarity. Middle, general structure of palladium reagents. RuPhos, 2-dicyclohexylphosphino-2',6'-diisopropoxybiphenyl. Right, the complexes were stored in screw-cap vials under air and weighed out on the bench top.

the rate of the palladium-mediated reaction was comparable to that of the maleimide ligation, where 70% of the products resulted from the reaction with palladium-tolyl complex (**1A-OTf**). Notably, the palladium-mediated conjugation outperformed the maleimide ligation at pH 5.5 where only the arylated product was formed.

The optimized conditions (0.1 M Tris buffer, 5% CH₃CN, pH 7.5, room temperature) were used for further evaluation of the substrate scope. Palladium complexes containing chloride, bromide and iodide counterions were all found to produce the desired product (Fig. 2, **1A-Cl**, **1A-Br** and **1A-I**). This method can be used to functionalize unprotected peptides with a variety of important groups, including fluorescent tags (**1C**, **1D**), affinity labels (**1E**), bioconjugation handles (aldehyde **1F**, ketone **1G** and alkyne **1H**) and photochemical cross-linkers (**1I**), as well as complex drug molecules (**1J**). Vinyl palladium complexes were also shown to be competent in this transformation

c, Model reaction with a peptide substrate (top) and the liquid chromatography mass spectrometry (LC-MS) trace of the crude reaction mixture after 5 min (bottom). The mass spectrum of the arylated product is shown in the inset. Sequence of peptide **P1**: NH₂-RSNFYLGACGLAHDKAT-C(O)NH₂. The reaction was quenched by the addition of 3-mercaptopropionic acid (3 equiv. to **1A-OTf**) before LC-MS analysis. At high reaction concentrations ($\geq 100 \mu\text{M}$) a cloudy precipitate formed after the addition of palladium reagent presumably owing to low solubility of the complex in the aqueous solvent. These reactions still produced the desired bioconjugate in high yields (Supplementary Information).

(Supplementary Information). Importantly, the palladium(II) complexes are stable under ambient conditions, and can be stored in closed vials under air at 4 °C for over four months. Long-term stability of **1A-I**, **1A-Br**, **1A-Cl** and **1A-OTf** was evaluated; only the complex bearing the trifluoromethanesulfonate counterion (**1A-OTf**) showed some degradation ($\leq 15\%$) after 20 weeks (Supplementary Information). Nevertheless, the 'aged' reagent still exhibited reactivity comparable to the freshly made complex.

The stability of our arylated peptides was compared to that of conjugates formed from reactions with reagents including *N*-ethyl maleimide, 2-bromoacetamide and benzyl bromide. The *S*-arylated peptide was shown to be stable towards acids, bases and external thiol nucleophiles (Supplementary Information). In contrast, the corresponding acetamide derivative was unstable under acidic and basic conditions, and the maleimide conjugate decomposed in the presence of base and

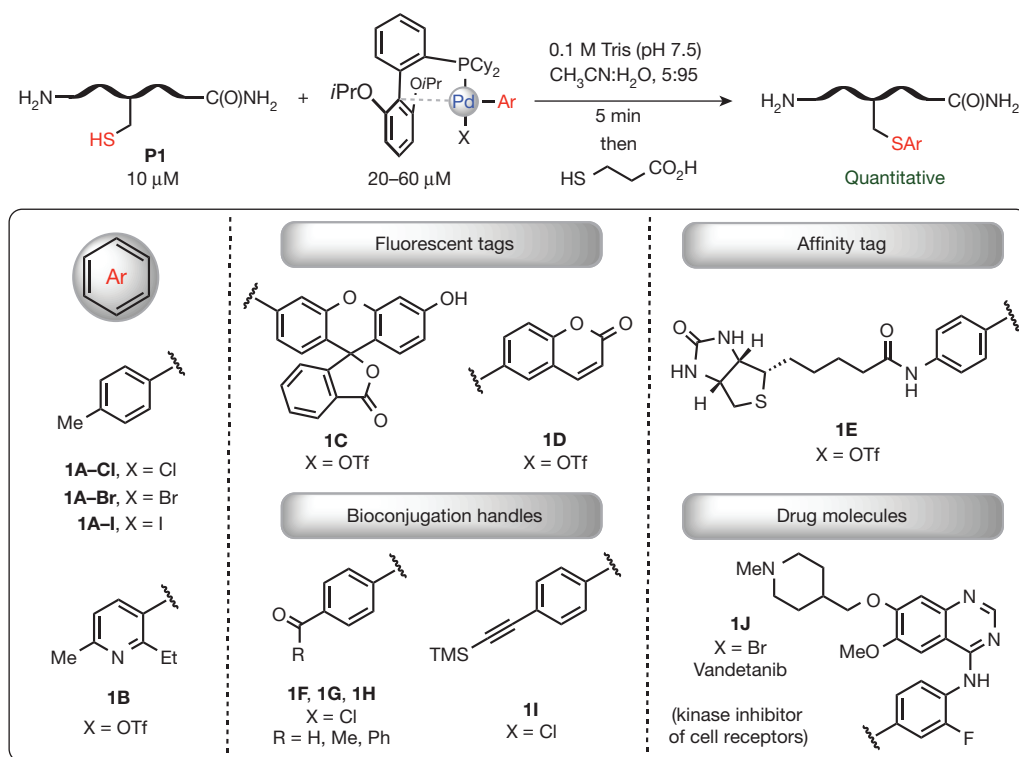


Figure 2 | The substrate scope of cysteine arylation using organometallic palladium reagents. Top, the reaction studied. Bottom, a selection of palladium reagents was used to test the effect of the leaving group (X) on the reactivity and explore the substrate scope with regard to biologically relevant

groups (fluorescent tags, bioconjugation handles, affinity tag and a drug molecule). Full conversion of starting peptide **P1** into the corresponding arylated products was observed in all the cases shown, as confirmed by LC-MS. For exact reaction procedures and conditions, see Supplementary Information.

exogenous thiol. Finally, comparable stability of both aryl and benzyl conjugates to treatment with the periodic acid oxidant at 37 °C was observed. However, additional tuning of the electronic properties of the aromatic ring of the arylated peptide could be achieved by installing an electron-withdrawing cyano group in the *para* position. This modification significantly decreased the amount of oxidation, producing the most stable peptides across all the evaluated conjugates.

Notably, installing the *para* cyano group in the benzyl conjugates had no effect on oxidation (Supplementary Information).

We further explored this reaction with proteins. Three antibody mimetic proteins²¹ (**P4–P6**; Fig. 3) were expressed that contained a cysteine at structurally distinct positions including the amino-terminus, the carboxy-terminus and a loop. The same proteins without cysteines were used as controls to confirm the selectivity of the reaction

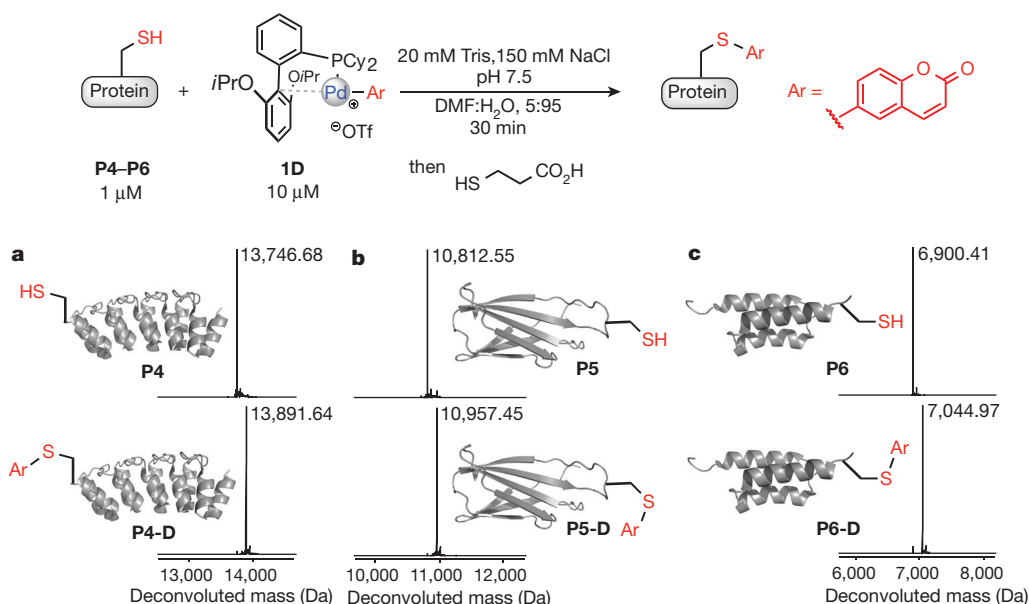


Figure 3 | Protein modification using the developed palladium reagents. Cysteine residues at the N terminus (**P4**; a), a loop (**P5**; b) and the C terminus (**P6**; c) of proteins are quantitatively modified with coumarin after the reaction with palladium complex **1D** (top). Deconvoluted mass spectra of the full

protein peaks are shown for the starting proteins (**P4–P6**) and reactions with coumarin-palladium complexes after 30 min (bottom). Three-dimensional structures of proteins **P4–P6** and the arylated products (**P4-D**, **P5-D** and **P6-D**) are presented next to the corresponding mass spectra.

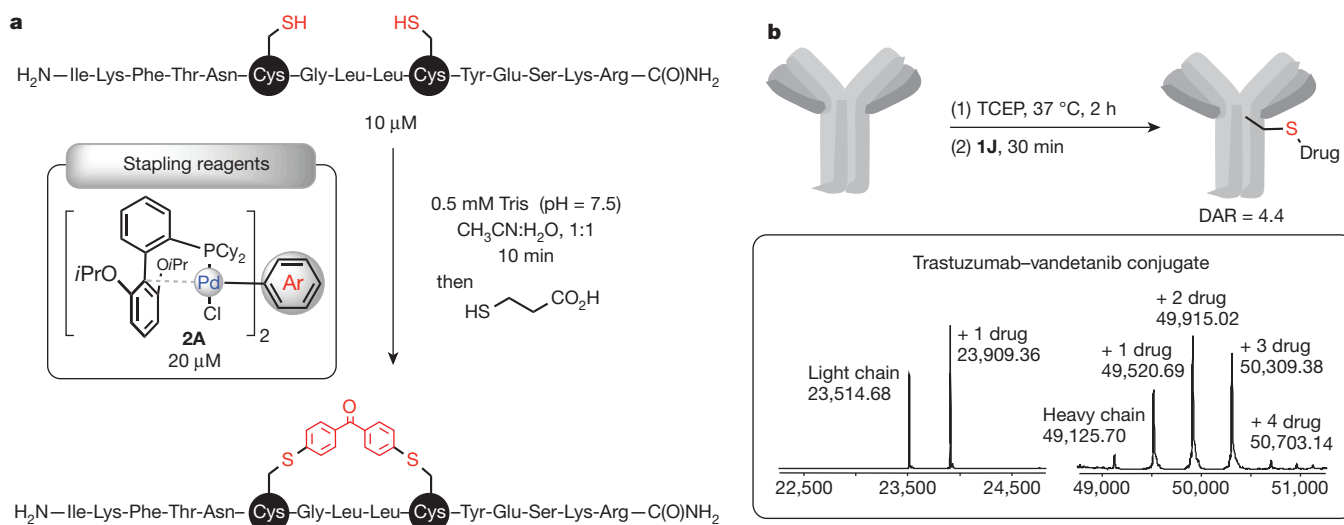


Figure 4 | Peptide stapling and antibody drug conjugate formation using palladium-based reagents. **a**, Stapling of a model peptide using bis-palladium reagent **2A**. Reaction conditions: **P3** (10 μ M), complex **2A** (20 μ M), Tris buffer (50 mM; pH 7.5), $\text{CH}_3\text{CN}:\text{H}_2\text{O}$ = 1:1, room temperature, 10 min. The reaction was quenched by the addition of 3-mercaptopropionic acid (6 equiv. to **2A**) before the LC-MS analysis. **b**, Synthesis of 'linker-free' antibody-drug

conjugates (ADCs). Trastuzumab antibody is represented as the two-tone grey structure at the top. Deconvoluted mass spectra of the fully reduced and deglycosylated ADCs are shown. Drug-to-antibody ratio (DAR) represents the average number of drugs per antibody. See Supplementary Information for details.

(P7–P9). All three proteins (P4–P6) were quantitatively tagged with either coumarin (Fig. 3) or a drug molecule (Supplementary Information) within 30 min at 1 μ M protein concentration. No arylated product was generated for proteins lacking a cysteine. The presence of small amounts of organic co-solvents is required for efficient bioconjugation, and lower product yields were observed when <5% of the co-solvent was used (Supplementary Information). The fast kinetics and high efficiency of the reactions at low micromolar and low nanomolar (Supplementary Information) protein concentrations are in contrast to reported bioconjugation methods using organometallic reagents, where longer reaction times were needed and generally lower conversions were observed^{3,22}.

The developed protocol was used to arylate an engineered cysteine residue in the C-terminal region of diphtheria toxin A-chain (DTA) fused to the lethal factor N-terminal domain (LF_N-DTA-Cys, see Supplementary Information)²³. The modified LF_N-DTA-Cys variant was readily separated from the remaining palladium species, ligands and other small molecules using commercially available size-exclusion chromatography columns (91% of palladium was removed, as determined by inductively coupled plasma mass spectrometry (ICP-MS) analysis of the purified protein sample, see Supplementary Information). The modified and purified LF_N-DTA-Cys variant displayed similar activity (half-maximal effective concentration EC_{50} = 0.40 ± 0.09 nM) in a cell-based protein synthesis inhibition assay compared to the control (a serine mutant LF_N-DTA-Ser showed EC_{50} = 0.25 ± 0.05 nM) (Supplementary Information).

Stapled peptides have shown significant promise as next-generation therapeutics^{24,25}. However, there are limited methods for the synthesis of these bioconjugates with structurally diverse linkers²⁶, which hinders the systematic investigation of the effect of the linker on the properties of the stapled peptides²⁷. We envisioned that palladium reagents containing two electrophilic metal centres could be efficiently used to cross-link two cysteine residues on a peptide chain, thereby providing access to stapled peptides with various aryl linkers (Fig. 4a). Indeed, running the reaction at 10 μ M concentration of peptide in a 1:1 (v/v) acetonitrile/water solvent mixture at pH 7.5 using a twofold excess of the bis-palladium complex **2A** resulted in quantitative formation of the target stapled peptide within 10 min (Fig. 3a and Supplementary Information). Considering the availability of commercially or other-

wise easily accessible diarylhalide reagents, this approach provides facile access to a diverse aryl-linker space for stapled peptides²⁸.

Antibody-drug conjugates (ADCs) are a promising class of biotherapeutics, which combine the potency of cytotoxic drugs with the target specificity of monoclonal antibodies²⁹. We aimed to attach drug molecules directly to cysteine residues in antibodies through the developed palladium conjugation chemistry. The drug payload vandetanib was used to form palladium complex **1J** (Fig. 2) by making use of the aryl bromide present in its structure. Treating partially reduced trastuzumab antibody³⁰ with **1J** readily produced ADCs with a 4.4 drug to antibody ratio (DAR; Fig. 4b). The purified arylated ADCs (94% of palladium was removed, as determined by ICP-MS analysis of the purified ADCs, see Supplementary Information) retained binding affinity (dissociation constant K_D = 0.3 ± 0.2 nM) to recombinant HER2 compared to the unmodified trastuzumab antibody (Supplementary Information). While traditional ADCs use various linkers to attach drug molecules to antibodies, our method significantly expands the structural space of ADCs by providing the capability to directly attach drug molecules containing native or pre-installed aryl halide or phenol functional groups. The therapeutic potential of this class of 'linker-free' ADCs will be investigated in the future.

We note that the ease of preparation, storage and use of the present palladium reagents make them particularly attractive for routine application in chemistry, biology, medicine and materials science. Further evolution of the metals and ligands employed will probably provide an extended set of organometallic bioconjugation reagents with altered selectivity and efficiency, allowing for functionalization of other amino acid residues.

Received 4 February; accepted 10 September 2015.

- Crabtree, R. H. in *The Organometallic Chemistry of the Transition Metals* 6th edn (Wiley, 2014).
- Diederich, F. & Stang, P. J. (eds) *Metal-catalyzed Cross-coupling Reactions* (Wiley & Sons, 2008).
- Antos, J. M. & Francis, M. B. Transition metal catalyzed methods for site-selective protein modification. *Curr. Opin. Chem. Biol.* **10**, 253–262 (2006).
- Yang, M., Li, J. & Chen, P. R. Transition metal-mediated bioorthogonal protein chemistry in living cells. *Chem. Soc. Rev.* **43**, 6511–6526 (2014).

5. Walsh, C. T., Garneau-Tsodikova, S. & Gatto, G. J. Protein posttranslational modifications: the chemistry of proteome diversifications. *Angew. Chem. Int. Edn* **44**, 7342–7372 (2005).
6. Rabuka, D. Chemoenzymatic methods for site-specific protein modification. *Curr. Opin. Chem. Biol.* **14**, 790–796 (2010).
7. Spicer, C. D. & Davis, B. G. Selective chemical protein modification. *Nature Commun.* **5**, 4740 (2014).
8. Chalker, J. M., Bernardes, G. J. L., Lin, Y. A. & Davis, B. G. Chemical modification of proteins at cysteine: opportunities in chemistry and biology. *Chem. Asian J.* **4**, 630–640 (2009).
9. Cal, P. M. S. D., Bernardes, G. J. L. & Gois, P. M. P. Cysteine-selective reactions for antibody conjugation. *Angew. Chem. Int. Edn* **53**, 10585–10587 (2014).
10. Toda, N., Asano, S. & Barbas, C. F. Rapid, stable, chemoselective labeling of thiols with Julia–Kocienski-like reagents: a serum-stable alternative to maleimide based protein conjugation. *Angew. Chem. Int. Edn* **52**, 12592–12596 (2013).
11. Lyon, R. P. *et al.* Self-hydrolyzing maleimides improve the stability and pharmacological properties of antibody–drug conjugates. *Nature Biotechnol.* **32**, 1059–1062 (2014).
12. Simmons, R. L., Yu, R. T. & Myers, A. G. Storable arylpalladium(II) reagents for alkene labeling in aqueous media. *J. Am. Chem. Soc.* **133**, 15870–15873 (2011).
13. Cheng, G., Lim, R. K. V., Li, N. & Lin, Q. Storable palladacycles for selective functionalization of alkyne-containing proteins. *Chem. Commun.* **49**, 6809–6811 (2013).
14. Barder, T. E., Biscoe, M. R. & Buchwald, S. L. Structural insights into active catalyst structures and oxidative addition to (biaryl)phosphine–palladium complexes via density functional theory and experimental studies. *Organometallics* **26**, 2183–2192 (2007).
15. Lee, C.-F., Liu, Y.-C. & Badsara, S. S. Transition-metal-catalyzed C–S bond coupling reaction. *Chem. Asian J.* **9**, 706–722 (2014).
16. Bong, D. T. & Ghadiri, M. R. Chemoselective Pd(O)-catalyzed peptide coupling in water. *Org. Lett.* **3**, 2509–2511 (2001).
17. Korneeva, E. N., Ovchinnikov, M. V. & Kostić, N. M. Peptide hydrolysis promoted by polynuclear and organometallic complexes of palladium(II) and platinum(II). *Inorg. Chim. Acta* **243**, 9–13 (1996).
18. Tilley, S. D. & Francis, M. B. Tyrosine-selective protein alkylation using π -allylpalladium complexes. *J. Am. Chem. Soc.* **128**, 1080–1081 (2006).
19. Shafer, D. E., Inman, J. K. & Lees, A. Reaction of tris(2-carboxyethyl)phosphine (TCEP) with maleimide and α -haloacyl groups: anomalous elution of TCEP by gel filtration. *Anal. Biochem.* **282**, 161–164 (2000).
20. Gorin, G., Martic, P. A. & Doughty, G. Kinetics of the reaction of *N*-ethylmaleimide with cysteine and some congeners. *Arch. Biochem. Biophys.* **115**, 593–597 (1966).
21. Gilbreth, R. N. & Koide, S. Structural insights for engineering binding proteins based on non-antibody scaffolds. *Curr. Opin. Struct. Biol.* **22**, 413–420 (2012).
22. Kung, K. K.-Y. *et al.* Cyclometalated gold(III) complexes for chemoselective cysteine modification via ligand controlled C–S bond-forming reductive elimination. *Chem. Commun.* **50**, 11899–11902 (2014).
23. Arora, N. & Leppla, S. H. Fusions of anthrax toxin lethal factor with shiga toxin and diphtheria toxin enzymatic domains are toxic to mammalian cells. *Infect. Immun.* **62**, 4955–4961 (1994).
24. Bird, G. H., Gavathiotis, E., LaBelle, J. L., Katz, S. G. & Walensky, L. D. Distinct BimBH3 (BimSAHB) stapled peptides for structural and cellular studies. *ACS Chem. Biol.* **9**, 831–837 (2014).
25. Verdine, G. L. & Hilinski, G. J. Stapled peptides for intracellular drug targets. *Methods Enzymol.* **503**, 3–33 (2012).
26. Lau, Y. H., de Andrade, P., Wu, Y. & Spring, D. R. Peptide stapling techniques based on different macrocyclization chemistries. *Chem. Soc. Rev.* **44**, 91–102 (2015).
27. Writing the macrocycle manual. *Nature Chem. Biol.* **10**, 693 (2014).
28. Spokoiny, A. M. *et al.* A perfluoroaryl-cysteine S_NAr chemistry approach to unprotected peptide stapling. *J. Am. Chem. Soc.* **135**, 5946–5949 (2013).
29. Chari, R. V. J., Miller, M. L. & Widdison, W. C. Antibody–drug conjugates: an emerging concept in cancer therapy. *Angew. Chem. Int. Edn* **53**, 3796–3827 (2014).
30. Sun, M. M. C. *et al.* Reduction-alkylation strategies for the modification of specific monoclonal antibody disulfides. *Bioconjug. Chem.* **16**, 1282–1290 (2005).

Supplementary Information is available in the online version of the paper.

Acknowledgements Financial support for this work was provided by the National Institutes of Health (GM-58160; GM-110535; postdoctoral fellowship for A.M.S., 1F32GM101762), the MIT start-up fund (B.L.P.), a Damon Runyon Cancer Research Foundation Award (B.L.P.) and the Sontag Foundation Distinguished Scientist Award (B.L.P.). C.Z. is the recipient of the George Büchi Research Fellowship and the Koch Graduate Fellowship in Cancer Research of MIT. We thank R. J. Collier (Harvard) for contributing select laboratory equipment used in this study. We thank the Biological Instrument Facility of MIT for providing the Octet BioLayer Interferometry System (NSF S10 OD016326). We are indebted to the NERCE facility (U54A1057159) for expressing the toxin proteins. We thank M. Lu and A. Rabideau for help with cell assays. The Varian 300 spectrometer used for portions of this work was purchased with funds from NSF (grant CHE-9808061). The departmental X-ray diffraction instrumentation was purchased with the help of funding from NSF (CHE-0946721). We are grateful to P. Müller (MIT) for X-ray crystallographic analysis of **1A-OTf-CH₃CN** and to A. Rancier (Merck) for the ICP-MS analysis.

Author Contributions S.L.B. conceived the idea of using palladium(II) reagents for bioconjugation; E.V.V., C.Z., A.M.S., B.L.P. and S.L.B. designed the research; E.V.V. and C.Z. conducted the majority of the experimental work; and A.M.S. conducted initial feasibility experiments. E.V.V. and C.Z. wrote the manuscript. All authors commented on the final draft of the manuscript and contributed to the analysis and interpretation of the data.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to B.L.P. (blp@mit.edu) and S.L.B. (sbuchwal@mit.edu).

Decadal slowdown of a land-terminating sector of the Greenland Ice Sheet despite warming

Andrew J. Tedstone¹, Peter W. Nienow¹, Noel Gourmelen¹, Amaury Dehecq^{1,2}, Daniel Goldberg¹ & Edward Hanna³

Ice flow along land-terminating margins of the Greenland Ice Sheet (GIS) varies considerably in response to fluctuating inputs of surface meltwater to the bed of the ice sheet. Such inputs lubricate the ice-bed interface, transiently speeding up the flow of ice^{1,2}. Greater melting results in faster ice motion during summer, but slower motion over the subsequent winter, owing to the evolution of an efficient drainage system that enables water to drain from regions of the ice-sheet bed that have a high basal water pressure^{2,3}. However, the impact of hydrodynamic coupling on ice motion over decadal timescales remains poorly constrained. Here we show that annual ice motion across an 8,000-km² land-terminating region of the west GIS margin, extending to 1,100 m above sea level, was 12% slower in 2007–14 compared with 1985–94, despite a 50% increase in surface meltwater production. Our findings suggest that, over these three decades, hydrodynamic coupling in this section of the ablation zone resulted in a net slowdown of ice motion (not a speed-up, as previously postulated¹). Increases in meltwater production from projected climate warming may therefore further reduce the motion of land-terminating margins of the GIS. Our findings suggest that these sectors of the ice sheet are more resilient to the dynamic impacts of enhanced meltwater production than previously thought.

The GIS is losing mass at an accelerating rate^{4,5}, as a result both of increased surface melting⁶ and of enhanced ice discharge from accelerating marine-terminating glaciers⁵. Enhanced melting accounts for about 60% of the GIS mass lost since 2000 (ref. 5); summer air temperatures over the southwest GIS warmed by 0.9 °C during 1994–2007 (ref. 7), and the amount of meltwater produced during the summers of 2007–12 (except 2009) is without precedent in the past 50 years of reanalysis-forced reconstructions⁸. During 1993–2012, the average annual melt doubled from that which occurred during 1961–90 (ref. 8).

While the acceleration of marine-terminating glaciers is believed to be driven primarily by processes operating at the ice–ocean interface, atmospheric forcing can also change ice motion at both land- and marine-terminating glaciers through the delivery of surface meltwater to the ice-sheet bed^{1,9}. Surface meltwaters can drain rapidly to the ice-sheet bed via moulins and supraglacial lake drainage events, both of which provide direct surface-to-bed connectivity and a mechanism by which surface meltwater could influence basal motion^{10–12}. It has been suggested that this mechanism could lead to a positive feedback between enhanced surface meltwater production and ice-sheet motion, as ice would move more quickly to lower elevations, where temperatures are warmer^{1,13}.

Other studies have highlighted the importance of the subglacial drainage system in controlling the relationship between surface melting and ice motion through changes in system capacity and morphology^{14–16}. During summer, rapid increases in meltwater from the ice-sheet surface result in periods in which the subglacial drainage system is more highly pressurized, leading to transient periods when the water pressure below exceeds the pressure caused by the weight of

the ice above, resulting in enhanced basal sliding¹⁵. However, the capacity of the subglacial drainage system then increases in response^{14,17}, introducing a negative feedback that lowers the water pressure and reduces basal sliding^{16,18}. By the end of summer, an efficient drainage system has evolved upglacier^{15,16}, draining surrounding regions of the ice-sheet bed that were previously hydraulically isolated. This reduces basal lubrication during the subsequent winter, counteracting the summer speed-up and making net annual ice motion relatively insensitive to summer melting^{2,3}.

Despite these advances in understanding the coupled hydrodynamics of ice-sheet flow, it remains unclear whether enhanced surface melting has a long-term impact on annual ice motion. Eight global positioning system (GPS) stations on a transect extending 130 km inland in the southwest GIS showed an average 10% decrease in ice flow from 1991 to 2007, during a period in which surface melt increased markedly, but there was considerable spatial variability¹⁹. The slowdown trend at lower sites continued into 2012 (ref. 20). Meanwhile, from 2009 to 2012 a small acceleration signal was observed above the altitude of the equilibrium line (where annual ice accumulation is equal to annual ablation), at ~1,500 metres above sea level (m.a.s.l.; ref. 21). The parametrization of basal lubrication in higher-order ice-sheet models, using observations from southwest Greenland, suggests that basal lubrication is unlikely to increase the contribution of the ice sheet to sea-level rise by more than 5% of the contribution that would be expected from a negative surface mass budget alone, and could conceivably act as a negative feedback upon ice motion²².

Here we present observations of annual GIS motion spanning three decades, which extend back to 1985. Our ~8,000-km² study area extends along about 170 km of the predominantly land-terminating margin of the west GIS, to ~50 km inland, and to ~1,100 m.a.s.l. (Fig. 1). We apply feature tracking (see Methods) to 475 pairs of remotely sensed optical Landsat images separated by approximately one year²³. Next, we derive robust ice-motion and uncertainty estimates over periods of approximately one to two years from 1985 to 2014 (Fig. 2b and Extended Data Fig. 4), and over multiyear reference periods spanning, first 1985–94, capturing the period before air temperatures began to warm⁷; and second 2007–14, corresponding to the recent series of summers with record melting⁸.

Ice motion shows a clear regional slowdown (Fig. 1), with 84% of the study area flowing more slowly in 2007–14 than in 1985–94 (Fig. 1a). On average, ice motion slowed by 12% across the study area. Slowdown was strongest (~15–20%) at elevations below around 800 m.a.s.l. (Fig. 1b). Isolated areas experienced speed-up in 2007–14 compared with 1985–94. In the far northeast, the speed-up can probably be attributed to the dynamics of the neighbouring marine-terminating Jakobshavn Isbrae, which, like many of Greenland's marine-terminating glaciers, has accelerated since the mid-1990s (ref. 5).

We can divide the ice-motion record (Fig. 2b) into two statistically significant periods (see Methods). Segmented linear regression ($R^2 = 0.79$) shows that there was no significant trend in ice motion during 1985–2002 ($P = 0.85$). The slowdown in motion probably

¹School of GeoSciences, University of Edinburgh, Edinburgh EH8 9XP, UK. ²Université Savoie Mont-Blanc, Polytech Annecy-Chambéry, LISTIC, BP 80439, 74944 Annecy-le-Vieux cedex, France.

³Department of Geography, University of Sheffield, Sheffield S10 2TN, UK.

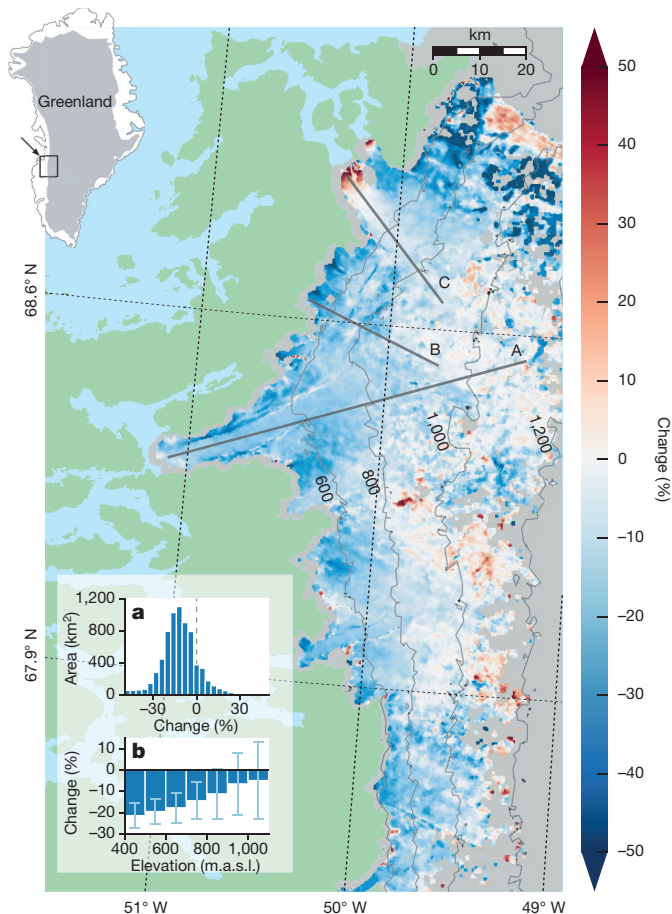


Figure 1 | Study area in the ablation zone of the western GIS. In the main figure, the colour scale shows the percentage change in ice velocities during the 2007–14 reference period compared with the 1985–94 reference period (see main text). Transects A, B and C correspond to data in Fig. 3. Ice surface contours (pale grey lines) are from ref. 30. Grey shading denotes areas where ice velocities cannot be resolved; green shading denotes land areas; light blue shading denotes inland and coastal waters. Inset: **a**, percentage changes in ice velocities in 4% bins; **b**, median percentage change in each 100-m elevation band between 400 m.a.s.l. and 1,100 m.a.s.l., $\pm 1\sigma$ (see Methods).

began around 2002, and, despite interannual variability, there was a robust overall trend of -1.5 m yr^{-2} during 2002–14 ($P < 0.01$). Meanwhile, we can divide surface meltwater production (Fig. 2a) into three statistically significant periods (see Methods): a sustained 'low' melt of 2.1 water equivalent (w.e.) m yr^{-1} during 1985–93; a rising melt during 1993–2002; and a sustained high melt of 3.2 w.e. m yr^{-1} during 2002–14, coincident with the slowdown in ice motion. Overall there was a 49.8% rise in surface meltwater production across our study area between 1985–94 and 2007–14.

We explored temporal variability in ice motion along three transects (Fig. 1), selected to represent different ice-marginal conditions. Transect A extends 80 km inland from the Nordenskjöld glacier, which has a lacustrine-terminating margin; transect B extends about 30 km inland from a land-terminating margin; and transect C extends about 30 km inland from the marine-terminating Alangordliup sermia. Transects A and B slowed down during 2000–14 to attain velocities, averaged along the transect, that were respectively 19% and 18% slower in 2013–14 than during 1985–94 (Fig. 3a, b). Ice-motion characteristics at the marine-terminating transect C were more complex (Fig. 3c). The transect slowed on average from the mid-2000s to 2014, although ice motion within 10 km of the margin sped up in the late 2000s following earlier slowdown, and by 2013–14 was flowing up to about 50 m yr^{-1} faster than during the 1985–94 reference period. Such behaviour is in line with other tidewater glaciers that have recently accelerated⁵.

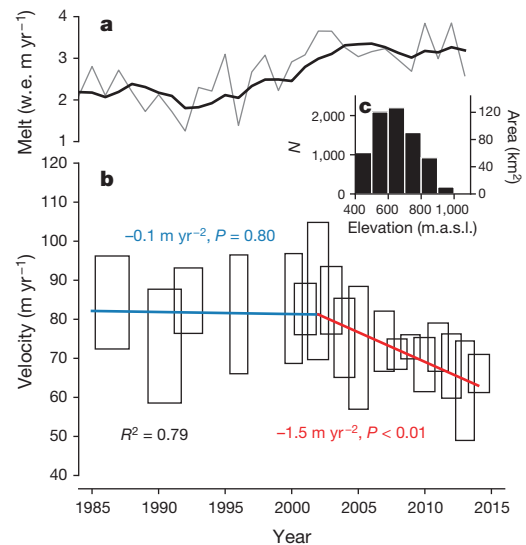


Figure 2 | Surface melting and ice motion averaged over the study area. **a**, Annual mean modelled surface melt (grey), smoothed with a five-year moving mean (black), both in water equivalent (w.e.) m per year (see Methods). **b**, Median ice velocities during each period (black boxes) calculated using the common sampling pixels across the time series, $\pm 1\sigma$ (see Methods). The width of each box corresponds to the total timespan of the pairs of Landsat images acquired during each period. The height of each box corresponds to $\pm 1\sigma$ (see Methods). Blue and red lines illustrate the trends in ice velocity computed by segmented linear regression weighted by 1σ . **c**, Altitudinal distribution of the common sampling pixels used to compute the velocities in **b**.

The slowdown signal across our predominantly land-terminating region extends up to about 1,100 m.a.s.l. (Fig. 1), where the mean ice thickness is roughly 850 m (ref. 24). The clear deceleration in ice motion requires a decrease in rates of either internal ice deformation, or basal motion, or both. Melting has caused marginal thinning of the GIS^{25–27}. During 1993–98, land-terminating glaciers on the west GIS margin thinned by $0.02\text{--}0.23 \text{ m yr}^{-1}$ below 1,000 m.a.s.l. (ref. 25). Our study area thinned by about 0.2 m yr^{-1} during 2003–07 (ref. 26) and this rate increased to $1\text{--}1.5 \text{ m yr}^{-1}$ during 2011–14 (ref. 27). We modelled the velocity change that would be caused by 10–20 m of ice thinning (and the associated gradient changes) along transect A over the 1985–2014 study period (see Methods), corresponding to a maximum thinning rate of about 0.6 m yr^{-1} . The resulting change in driving stress can explain only around 17–33% of the observed overall 12% slowdown signal beyond 10 km from the ice-sheet margin, and can explain none of the slowdown beyond 50 km from the margin (Extended Data Fig. 5c). Thus, while a component of the observed slowdown can be explained by changes in driving stress through ice thinning, the majority of the slowdown (that is, the remaining 67–83%) must be the result of processes operating at the ice-bed interface that cause a reduction in basal motion.

Previous studies have suggested that the coupling between surface melting and basal motion is self-regulating, such that there is no statistically significant relationship between melting and ice motion over annual timescales^{2,20}. In agreement with these studies, we find no relationship between annual melt volume and annual ice motion ($R^2 = 0.08$). There is, however, a significant relationship between antecedent melt volumes and ice motion (Extended Data Table 1). The mean melt volume from each observation period and the previous year combined explain 23% of ice motion ($P < 0.05$), increasing to 44% when the previous four years of melt are included. Moreover, melt volumes explain 50% of ice motion when the mean melt volume is calculated using only the previous three years' data ($P < 0.01$).

We therefore suggest that sustained high production of surface meltwater (Fig. 2a) is responsible for the slowdown. Observations from

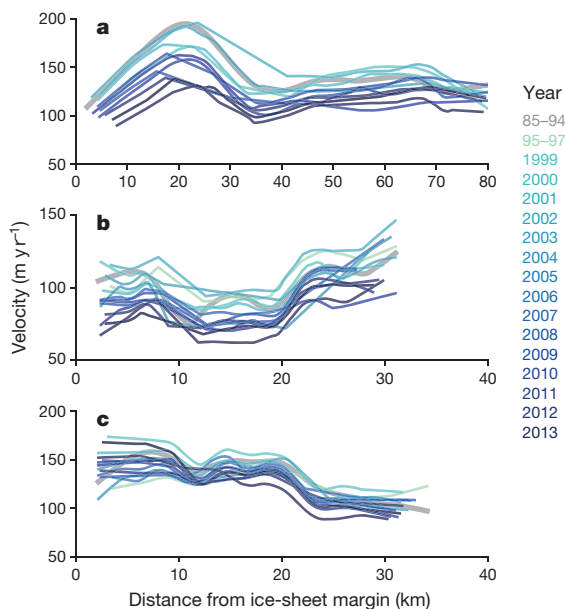


Figure 3 | Ice velocities along three transects in the study area. Results for **a**, transect A; **b**, transect B; **c**, transect C. These transects are shown in Fig. 1. Only periods in which ice velocities are observed along at least 60% of each transect are shown. Velocities during the 1985–94 reference period are also shown.

the GIS show that during the melt season, the large cumulative increase in the rate of meltwater supply to the ice-sheet bed results in the expansion of a channelized subglacial drainage system¹⁶, even beneath ice that is around 1 km thick¹¹. As air temperatures warm, more meltwater at higher elevations allows an efficient drainage system to evolve, which extends further into the ice sheet. Summers of extreme melt result in a higher-capacity, more-extensive channelized drainage system, which therefore stays open at atmospheric pressure for longer after the melt ceases^{2,3}. Dye tracing of alpine glaciers²⁸ indicates that water transit speeds through unchannelized drainage systems are $\sim 0.01 \text{ m s}^{-1}$ ($\sim 850 \text{ m d}^{-1}$). Thus, while channels at atmospheric pressure beneath ice that is about 1 km thick close within hours to days¹¹, channels that stay open for longer, for example for just two days as opposed to one, have the capacity to evacuate notably more water from surrounding linked but unchannelized regions of the ice-sheet bed, causing more widespread dewatering of the ice–bed interface. Sources of meltwater, including frictional melting by basal slip and geothermal heat, will enable the water pressure to recover gradually through the subsequent winter, but may be insufficient to replace the stored waters evacuated during the previous melt season.

Previous observations have illustrated the importance of changes in the connectivity between channelized and unchannelized regions of the ice-sheet bed in controlling ice velocities late in the melt season¹². We postulate that unchannelized drainage regions and their connectivity to the channelized drainage system govern ice motion not only late in the melt season, but also during the following winter and spring. We suggest that, if increases in drainage efficiency occur year-on-year, gradual net drainage of water stored in unchannelized regions of the ice-sheet bed will result in reduced basal lubrication and net ice slowdown. Additional field observations—such as borehole arrays, transverse to subglacial channels, recording water pressure gradients (see, for example, ref. 29), together with hydrological modelling (see, for example, ref. 14)—are required to test the robustness of our hypothesis. Furthermore, while melt-driven seasonal evolution in subglacial drainage can affect the flow of tidewater glaciers⁹, the ongoing acceleration of these glaciers⁵ during a period of warming, in contrast to our observations, suggests that other processes are controlling their dynamics.

Our observations of GIS ice motion from three decades provide conclusive evidence that a 50% rise in meltwater production has not led to ice speed-up along a land-terminating margin; instead, average annual ice motion slowed by more than 15% at elevations below 800 m.a.s.l., and probably by at least 5% at elevations up to 1,100 m.a.s.l. Only about 17–33% of the slowdown can be explained by reduced internal deformation caused by ice thinning, and we therefore propose that, since 2002, increases in subglacial drainage efficiency associated with sustained larger melt volumes have reduced basal lubrication, resulting in slower ice flow. It remains unclear whether the observed slowdown occurs at elevations above 1,100 m.a.s.l., and whether the slowdown will migrate inland as enhanced melting extends to higher elevations and allows a more extensive efficient subglacial drainage system to evolve. Furthermore, our findings relate to land-terminating margins, but the forcing mechanisms that have driven the recent speed-up of many tidewater glaciers remain poorly understood^{5,26} and require a similar examination of annual ice motion over decadal timescales.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 2 April; accepted 1 September 2015.

- Zwally, H. J., Abdalati, W. & Herring, T. Surface melt-induced acceleration of Greenland ice-sheet flow. *Science* **297**, 218–222 (2002).
- Sole, A. *et al.* Winter motion mediates dynamic response of the Greenland ice sheet to warmer summers. *Geophys. Res. Lett.* **40**, 3940–3944 (2013).
- Tedstone, A. J. *et al.* Greenland ice sheet motion insensitive to exceptional meltwater forcing. *Proc. Natl Acad. Sci. USA* **110**, 19719–19724 (2013).
- Shepherd, A. *et al.* A reconciled estimate of ice-sheet mass balance. *Science* **338**, 1183–1189 (2012).
- Enderlin, E. M. *et al.* An improved mass budget for the Greenland ice sheet. *Geophys. Res. Lett.* **41**, 866–872 (2014).
- Hanna, E. *et al.* Ice sheet mass balance and climate change. *Nature* **498**, 51–59 (2013).
- Box, J. E., Yang, L., Bromwich, D. H. & Bai, L.-S. Greenland ice sheet surface air temperature variability: 1840–2007. *J. Clim.* **22**, 4029–4049 (2009).
- Fettweis, X. *et al.* Important role of the mid-tropospheric atmospheric circulation in the recent surface melt increase over the Greenland ice sheet. *Cryosphere* **7**, 241–248 (2013).
- Howat, I. M., Box, J. E., Ahn, Y., Herrington, A. & McFadden, E. M. Seasonal variability in the dynamics of marine-terminating outlet glaciers in Greenland. *J. Glaciol.* **56**, 601–613 (2010).
- Das, S. B. *et al.* Fracture propagation to the base of the Greenland Ice Sheet during supraglacial lake drainage. *Science* **320**, 778–81 (2008).
- Chandler, D. M. *et al.* Evolution of the subglacial drainage system beneath the Greenland Ice Sheet revealed by tracers. *Nature Geosci.* **6**, 195–198 (2013).
- Andrews, L. C. *et al.* Direct observations of evolving subglacial drainage beneath the Greenland ice sheet. *Nature* **514**, 80–83 (2014).
- Parizek, B. R. & Alley, R. B. Implications of increased Greenland surface melt under global-warming scenarios: ice-sheet simulations. *Quat. Sci. Rev.* **23**, 1013–1027 (2004).
- Schoof, C. Ice-sheet acceleration driven by melt supply variability. *Nature* **468**, 803–806 (2010).
- Bartholomew, I. *et al.* Seasonal variations in Greenland Ice Sheet motion: Inland extent and behaviour at higher elevations. *Earth Planet. Sci. Lett.* **307**, 271–278 (2011).
- Cowton, T. *et al.* Evolution of drainage system morphology at a land-terminating Greenland outlet glacier. *J. Geophys. Res.* **118**, 29–41 (2013).
- Röthlisberger, H. Water pressure in intra- and subglacial channels. *J. Glaciol.* **11**, 177–203 (1972).
- Hoffman, M. & Price, S. Feedbacks between coupled subglacial hydrology and glacier dynamics. *J. Geophys. Res. Earth Surf.* **119**, 414–436 (2014).
- van de Wal, R. S. W. *et al.* Large and rapid melt-induced velocity changes in the ablation zone of the Greenland ice sheet. *Science* **321**, 111–113 (2008).
- van de Wal, R. S. W. *et al.* Self-regulation of ice flow varies across the ablation area in south-west Greenland. *Cryosphere* **9**, 603–611 (2015).
- Doyle, S. H. *et al.* Persistent flow acceleration within the interior of the Greenland ice sheet. *Geophys. Res. Lett.* **41**, 899–905 (2014).
- Shannon, S. R. *et al.* Enhanced basal lubrication and the contribution of the Greenland ice sheet to future sea-level rise. *Proc. Natl Acad. Sci. USA* **110**, 14156–14161 (2013).
- Dehecq, A., Gourmelen, N. & Trounev, E. Deriving large scale glacier velocities from a complete satellite archive: application to the Pamir-Karakoram-Himalaya. *Remote Sens. Environ.* **162**, 55–66 (2015).
- Morlighem, M., Rignot, E., Mouginot, J., Seroussi, H. & Larour, E. Deeply incised submarine glacial valleys beneath the Greenland ice sheet. *Nature Geosci.* **7**, 418–422 (2014).

25. Sole, A., Payne, T., Bamber, J., Nienow, P. & Krabill, W. Testing hypotheses of the cause of peripheral thinning of the Greenland ice sheet: is land-terminating ice thinning at anomalously high rates? *Cryosphere* **2**, 205–218 (2008).
26. Pritchard, H. D., Arthern, R. J. & Vaughan, D. G. and Edwards, L. A. Extensive dynamic thinning on the margins of the Greenland and Antarctic ice sheets. *Nature* **461**, 971–975 (2009).
27. Helm, V., Humbert, A. & Miller, H. Elevation and elevation change of Greenland and Antarctica derived from CryoSat-2. *Cryosphere* **8**, 1539–1559 (2014).
28. Richards, K. *et al.* An integrated approach to modelling hydrology and water quality in glacierized catchments. *Hydrol. Processes* **10**, 479–508 (1996).
29. Hubbard, B., Sharp, M., Willis, I., Nielsen, M. & Smart, C. Borehole water-level variations and the structure of the subglacial hydrological system of Haut Glacier d'Arolla, Valais, Switzerland. *J. Glaciol.* **41**, 572–583 (1995).
30. Howat, I. M., Negrete, A. & Smith, B. E. The Greenland Ice Mapping Project (GIMP) land classification and surface elevation data sets. *Cryosphere* **8**, 1509–1518 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements A.J.T. acknowledges UK Natural Environment Research Council (NERC) studentships NE/152830X/1 and NE/J500021/1, a Scottish Alliance for

Geoscience, Environment and Society (SAGES) Postdoctoral/Early Career Researcher Exchange (PECRE) award, and a University of Edinburgh GeoSciences Moss scholarship. N.G. acknowledges European Space Agency Dragon 3 grant 10302, the Centre National d'Etudes Spatiales Tosca CESTENG project, and a fellowship from the Centre National d'Etudes Spatiales to A.D. This work made use of the resources provided by the Edinburgh Compute and Data Facility (ECDF) (<http://www.ecdf.ed.ac.uk/>). We thank P. Huybrechts for his work on the runoff/retention model used in this study. The Landsat imagery was provided by the United States Geological Survey and the European Space Agency third party missions program.

Author Contributions A.J.T., P.W.N. and N.G. designed this study. A.D., N.G. and A.J.T. developed the processing chain used for feature tracking of Landsat imagery. A.J.T., A.D. and N.G. processed the Landsat imagery. A.J.T. and D.G. calculated the impact of changing ice geometry upon ice motion. E.H. processed the melt data. A.J.T., N.G. and P.W.N. analysed the results. A.J.T., P.W.N. and N.G. wrote the manuscript. All authors discussed the results and edited the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.J.T. (a.j.tedstone@ed.ac.uk).

METHODS

Remote sensing of ice motion. We applied feature-tracking techniques to extract ice motion from Landsat Program imagery. Landsat images were obtained from the US Geological Survey (via the Earth Explorer catalogue at <http://earthexplorer.usgs.gov>) and the European Space Agency (via the Earth Observation Link catalogue at <https://earth.esa.int/web/guest/eoli>). Here we provide an outline of the processing strategy and the specific parameters that we used. A detailed description of the processing strategy is available elsewhere²³. Our approach builds individual annual velocity fields from feature tracking of Landsat pairs, overlapping in time and space. These velocity fields are then combined over interannual time periods in order to increase the robustness of the velocity estimates and to enable statistical determination of uncertainties.

We used images from the Landsat 5, 7 and 8 missions; the quality and quantity of images from Landsat missions 1–3 were insufficient to permit their use. We identified 475 image pairs with temporal baselines between 352 and 400 days, acquired from April to October over the 1985–2014 study period (see Supplementary Information). The temporal baseline of ~1 year was chosen to minimize the impact of seasonal flow variability upon interannual trends in ice velocity, as we are specifically interested in long-term changes in ice motion.

To enhance the images before feature tracking, we used principal component analysis to combine the optimum spectral bands, identified during testing as bands 2 and 3 for each satellite mission. A high-pass filter (using Sobel kernels) was used to compute the intensity gradients of each image, enhancing surface features such as crevasses and reducing the impact of features related to basal topography, which by definition are temporally stable. We then used the gradients as features to be tracked. The tracking (used to extract ice displacement) was performed on matching windows of 44 pixels (1,320 m) and a grid spacing of 8 pixels (240 m), while the search window was set automatically to correspond to the maximum expected displacement over the baseline duration between the two images, based on previous velocity observations³¹.

The processing strategy exploits the redundancy offered by multiple, spatio-temporally overlapping pairs to remove velocity outliers efficiently and produce robust velocity fields. First, we filter low-quality velocity estimates by applying a threshold to the signal-to-noise ratio returned during the feature tracking. The threshold value was determined by examining all velocity pairs to identify the value beyond which the median absolute deviation (MAD) of stable area velocities becomes asymptotic. We use a median-based approach to minimize the impact of outliers, and because the distribution of velocity tends not to follow a normal distribution²³. Next, for the period 2000–14, we group velocities into one-year time periods and for the period 1985–2000, we group velocities into two-year time periods (because there were fewer Landsat pairs available). This provides spatio-temporal redundancy in the velocity estimates at each pixel and enables us to quantify uncertainties. To produce the final velocity field for each period, we compute the median of all the available velocity estimates at each pixel. Lists of the Landsat pairs that contribute to each period are available in the Supplementary Information. Finally, we compute the 1σ uncertainty of the velocity estimate at each pixel in each period. To do so we fit a law of the form:

$$\sigma = \frac{k \text{MAD}}{2 N^{\alpha}} \quad (1)$$

where N is the number of velocities used to compute the median velocity, σ is the 1σ confidence interval, and k and α are the parameters to be determined. k and α were determined for each time period from the velocity estimates made over stable land areas whose true value is known and equal to 0. Uncertainty of the final ice velocity of each pixel at each time period is then obtained by extrapolating this relationship to on-ice areas with the appropriate values of MAD and N for a given pixel. Considering N and MAD at the pixel level allows the surface characteristics at the location of each pixel (for example, surface conditions or variability of velocity during the time period considered) to be taken into account. Pixels with $\sigma > 60 \text{ m yr}^{-1}$ are discarded in the subsequent analysis.

We computed the percentage change in ice velocities between 1985 and 1994, and between 2007 and 2014, at all of the pixels that are common to both periods (Fig. 1); we then computed the median percentage change, both over the whole study area and in elevation bands of 100 m a.s.l.. For each elevation band in Fig. 1b, we calculated the uncertainty of the percentage change by first estimating the uncertainty of the 1985–94 (e_1) and 2007–14 (e_2) velocities separately as $\sqrt{\sum_{i=1}^N \sigma_i^2 / N}$, and then computing $\sqrt{e_1^2 + e_2^2}$. Residual striping patterns in Fig. 1 are caused by lines of missing data in the Landsat 5 imagery.

To compute interannual median velocities (Fig. 2b), we discarded first periods in which less than 30% of the study area has observations. Then, the pixels common to all the retained periods (Extended Data Fig. 4) were selected so as to avoid temporal variation caused by spatial bias. For each period we calculated the med-

ian of the 8,025 temporally common pixels and the associated uncertainty σ , estimated as $\sqrt{\sum_{i=1}^N \sigma_i^2 / N}$. The altitudinal distribution of the common sampling pixels is shown in Fig. 2c.

The decision to discard periods in which less than 30% of the study area has observations is a compromise between calculating the median velocity of each period using the greatest possible number of pixels common to all periods, versus retaining the maximum possible temporal resolution. We examined the sensitivity of the ice-motion time series (Fig. 2b) to 40% and 50% thresholds. At 40%, there are 5,970 more sampling pixels in common than at 30%, but temporal resolution decreases as ice velocities observed in 2003–06 are no longer retained. The R^2 of the two-trend model decreases to 0.65. The rate of slowdown during 2002–14 increases from 1.5 m yr^{-2} to 1.9 m yr^{-2} ($P < 0.01$). At 50%, there are 8,397 more sampling pixels in common than at 30%. Velocities observed during the 1995–97 period are also discarded. The R^2 of the two-trend model is 0.63. The rate of slowdown during 2002–14 remains the same as at 40% ($P < 0.01$). There is no statistically significant trend in ice motion from 1985 to 2002 ($P > 0.05$) at either of the tested thresholds. From this sensitivity analysis, we conclude that our 30% threshold case yields the highest temporal resolution and also the most conservative trend in ice motion during 2002 to 2014.

Identification of trends in melting and ice motion. For each time series, we test whether it can be divided into temporally distinct populations separated by break dates. We apply the Mann–Whitney–Wilcoxon (MWW) rank sum test to investigate the effect of prescribing different break dates. We chose the MWW test over the t -test as we do not know whether the data set follows a normal distribution. The test computes the probability that the populations, separated by prescribed break dates, are similar.

The melt time series (Fig. 2a) consists of several years of sustained low melt, followed by a period of rising melt and then several years of sustained higher melt. We therefore test whether the melt time series can be split into three statistically different populations, with two break dates separating the periods of sustained low, rising and sustained high melt. We find that the break-date combinations of (i) 1992 and 2001, (ii) 1993 and 2001, and (iii) 1993 and 2002 are all significant with 95% confidence (Extended Data Fig. 2). To find the best possible combination of break dates, we compute the root-mean-squared error, or residuals, of the best-fitting three-trend segmented linear regression model (Extended Data Fig. 2). We observe the lowest residuals for break dates of 1991–93 and 2001–04. The combination of break dates that satisfies the MWW test and has the lowest residuals is 1993 and 2002. With this chosen combination of break dates, the probability that (i) 1985–93 is similar to 2002–13 is 0.02%; (ii) 1985–93 is similar to 1994–2001 is 2%; and (iii) 1994–2001 is similar to 2002–13 is 3%.

For the ice-motion time series, we first test whether it can be divided into two temporally distinct populations, in order to justify the use of segmented linear regression (Fig. 2). We apply the MWW test as previously. We find that for break dates beyond mid-2001, the null hypothesis (equal median) can be rejected (Extended Data Fig. 3a), meaning that the pre- and post-2001 populations are statistically different with 95% confidence. To test the ability of two distinct periods separated by a given break date to represent the velocity time series, we then compute the residuals of the best fitting two-trend segmented linear regression model for a set of break dates spanning the time period of the data set (Extended Data Fig. 3b). We observe a minimum for a break date in 2002, but with a region of low residuals spanning the period 1998 to 2004. We conclude from these tests that there are two distinct temporal populations of ice motion in our data set and that the break point occurs during the period 1998 to 2004. For our analysis (Fig. 2) we select a break date of 2002, which corresponds both with the lowest residuals and with the MWW test that suggests that the pre- and post-2002 populations are statistically different.

We then examine whether there is a statistically significant relationship between meltwater production and ice motion by applying linear regression analysis to investigate the extent to which variability in ice motion can be explained by, first, temporally coincident, and second, antecedent meltwater production (Extended Data Table 1). We quantify antecedent meltwater production in two different ways. In one scenario we calculate the mean during the period of observed ice motion and the preceding N years. In the other scenario we calculate the mean of only the preceding N years.

Impact of varying baseline durations on annual velocity. Images separated by baseline durations of 352–400 days were paired together for feature tracking. Here we examine the impact that the variable baseline duration has on the velocity field of each period.

The average start dates of the pairs that comprise each period are shown in Extended Data Fig. 1a. The start day becomes less variable once imagery from Landsat 7 comes online in 1999. The average baseline duration increases by about 15 days after 1999 (Extended Data Fig. 1b), increasing the proportion of the

baseline that is attributable to summer motion (defined as 1 May to 31 August, in common with previous studies^{2,3}) by about 2% (Extended Data Fig. 1c).

We use mean summer and winter velocities from Leverett Glacier sites S1 to S4 during the period 2009 to 2012 (refs 2, 3) to test the sensitivity of annual velocities to the varying baseline duration. Winter days in each period are ascribed velocities of 81.6 m yr^{-1} and summer days are ascribed velocities of 127.6 m yr^{-1} . We then estimate the mean annual velocity that would be expected for each period (Extended Data Fig. 1d). Variations in the baseline duration between periods are estimated to affect extracted annual ice motion by more than 2 m yr^{-1} . Furthermore, according to this analysis, increased baseline durations during the 2000s leads to a small artificial increase in ice motion caused by the feature-tracking method. This is in the opposite direction to the interannual slowdown signal that we observe in our study area, leading us to conclude that our slowdown trend is robust to varying baseline durations.

Impact of changing ice geometry on velocity. During the past 30 years, the GIS has thinned along its margins^{26,27,32}, changing the geometry of the ice mass. It is likely that these changes will have affected ice velocity by modifying driving stress. Here we evaluate the impact that thinning may have had on velocity along transect A (Fig. 1), in order to bound the extent to which our observed slowdown could be the result of geometric changes.

We characterize the surface velocity u_s as the sum of a basal sliding (u_b) and a vertical shear deformation (u_d) contribution³³:

$$u_s = u_b + u_d = C_b(\rho_i g S H)^m + \frac{A}{4}(\rho_i g S)^3 H^4 \quad (2)$$

where A is a temperature-dependent Glen's flow parameter, ρ_i is the density of ice, g is gravitational acceleration, S is the surface slope (positive where the surface lowers towards the margin), H is the ice thickness, and C_b and m are parameters related to basal sliding. A , C_b and m are in general poorly constrained, and are likely to vary spatially; however, the only assumptions we make in our analysis regarding these parameters are that they do not change markedly over the time interval of interest, and that m is less than or equal to 3. Note that our model allows for either the power-law rheological model of Weertman³⁴ or the Newtonian till model of Alley *et al.*³⁵. Thus the maximum deceleration predicted by the above model bounds the slowdown that can be explained by geometric changes alone. Below, we estimate this maximum deceleration to first order.

We introduce the variable λ , which represents the fraction of surface velocity that can be explained by vertical shear, that is, $u_d = \lambda u_s$, $u_b = (1 - \lambda)u_s$.

If we consider a small change δS in slope ($\delta S \ll S$), and a small change δH in ice thickness ($\delta H \ll H$), equation (2) leads to the following change in u_s :

$$\begin{aligned} \delta u_s &= \left(\frac{m u_b + 3 u_d}{u_s} \right) \frac{\delta S}{S} + \left(\frac{m u_b + 4 u_d}{u_s} \right) \frac{\delta H}{H} + \theta(\delta H^2, \delta S^2) \\ &= (m(1 - \lambda) + 3\lambda) \frac{\delta S}{S} + (m(1 - \lambda) + 4\lambda) \frac{\delta H}{H} + \theta(\delta H^2, \delta S^2) \end{aligned} \quad (3)$$

where the θ -notation is used to signify terms that are of order δH^2 and δS^2 or higher and thus are negligibly small. Again, we make no assumption regarding the spatial variability of λ , other than that it is between 0 and 1, as our aim is to find the

conditions under which velocity is most sensitive to thinning. With $\left(\frac{\delta S}{S} \right)$ positive

and $\left(\frac{\delta H}{H} \right)$ negative in equation (3), then at any point along the transect, and for any $m \leq 3$, the change to u_s cannot be more negative than when the flow is due to vertical shear, that is, when λ is equal to 1. The first-order relative change in surface velocity, $\frac{\delta u_s}{u_s}$, is thus bounded by:

$$3 \frac{\delta S}{S} + 4 \frac{\delta H}{H} \quad (4)$$

We use equation (4) to estimate the maximal impact of these thinning scenarios on ice velocity, solving every 240 m along transect A. To estimate the total ice thinning during 1985–2014, $\delta H = 10, 20$ are prescribed at the ice-sheet margin and linearly

interpolated along the transect to $\delta H = 0 \text{ m}$ at 100 km inland (equivalent to the equilibrium line altitude, $\sim 1,500 \text{ m a.s.l.}$, ref. 36) (Extended Data Fig. 5a). We add δH to current ice thickness along the transect²⁴ to set H to values appropriate for 1985. We prescribe the initial slope S as the mean slope in our study area, 0.02 mm^{-1} . The change in slope, δS , is calculated from the prescribed linear change in ice thickness over distance inland.

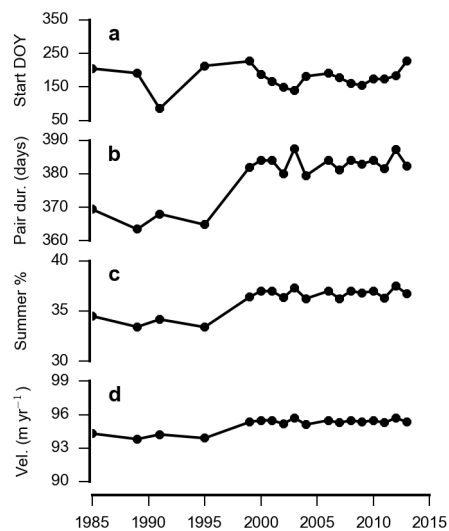
In Extended Data Fig. 5b we plot equation (4) according to the two thinning scenarios described above. These profiles represent the largest (that is, the most negative) percentage changes in velocity associated with the prescribed geometric change. We then convert these profiles to the predicted reductions in velocity and remove them from the observed 1985–94 velocities (Extended Data Fig. 5c), generating a lower bound for the 2007–14 velocities under the assumption that the observed slowdown was geometrically induced. Between about 0 km and 5 km from the ice margin for the $\delta H = 20 \text{ m}$ scenario, the observed slowdown is within the range predicted by equation (4), and we cannot reject the possibility that the thinning here was responsible for the slowdown. However, between 10–50 km inland, at most 17–33% of the observed slowdown can be attributed to changing ice-sheet geometry depending on the prescribed δH . By 60 km inland, there is essentially no net change in ice velocity attributable to geometrical changes.

Finally we must consider our assumption that the Glen's law parameter A can be treated as constant in time. Phillips *et al.*³⁷ showed that latent heat transferred to the ice from surface melt could warm glacial ice at depth, thereby leading to an increase in A . However, this process would result in acceleration rather than deceleration. Thus our decision not to consider temporal changes in A in our analysis of maximal geometrically induced slowdown is justified. We conclude that the slowdown that we have observed during 1985–2014 is not explicable by geometrical changes to the ice sheet alone, and instead must be dominated, at distances greater than $\sim 5 \text{ km}$ from the margin, by other processes that affect basal motion.

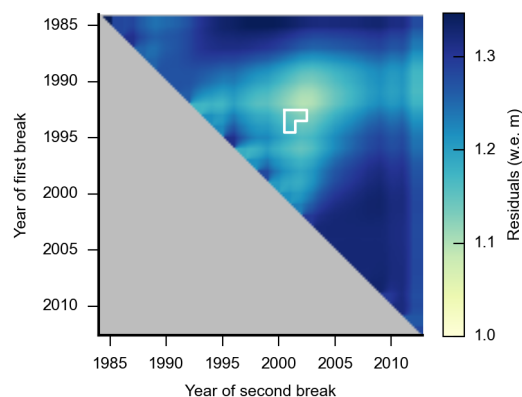
Surface melting. GIS annual melting was output from a runoff/retention model applied to downscaled ERA-interim data from the European Centre for Medium-Range Weather Forecasts (ECMWF) on an equal-area $5 \text{ km} \times 5 \text{ km}$ polar stereographic grid for the Greenland region³⁸. We calculated mean annual melt rates for the study area (67.45° N , 51.5° W to 69.2° N , 49.2° W). Interannual fluctuations and trends from several independent melt models show good agreement³⁹, including with the methodology used in this study.

Code availability. The feature-tracking algorithm is proprietary software developed and licensed by GAMMA Remote Sensing. Requests for the code underlying the processing strategy used here should be sent to the original authors²³. Similarly, requests for the code constituting the runoff/retention model should be addressed to its original authors⁴⁰.

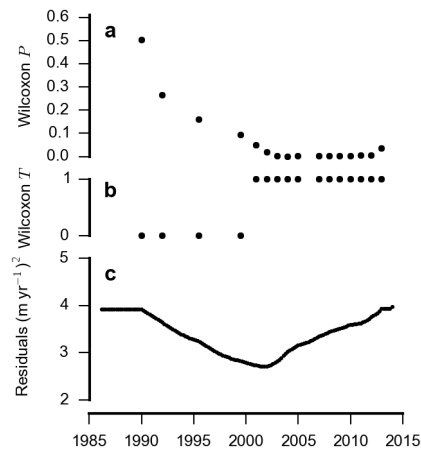
31. Joughin, I., Smith, B. E., Howat, I. M., Scambos, T. & Moon, T. Greenland flow variability from ice-sheet-wide velocity mapping. *J. Glaciol.* **56**, 415–430 (2010).
32. Krabill, W. *et al.* Greenland ice sheet: high-elevation balance and peripheral thinning. *Science* **289**, 428–430 (2000).
33. Cuffey, K. & Paterson, W. S. B. *The Physics of Glaciers* 3rd edn (Butterworth-Heinemann, 2010).
34. Weertman, J. On the sliding of glaciers. *J. Glaciol.* **3**, 33–38 (1957).
35. Alley, R. B., Blankenship, D. D., Rooney, S. T. & Bentley, C. R. Till beneath ice stream B: 4. A coupled ice-till flow model. *J. Geophys. Res. Solid Earth* **92**, 8931–8940 (1987).
36. van de Wal, R. S. W. *et al.* Twenty-one years of mass balance observations along the K-transect, West Greenland. *Earth Syst. Sci. Data* **4**, 31–35 (2012).
37. Phillips, T., Rajaram, H. & Steffen, K. Cryo-hydrologic warming: a potential mechanism for rapid thermal response of ice sheets. *Geophys. Res. Lett.* **37**, L20503 (2010).
38. Hanna, E. *et al.* Greenland ice sheet surface mass balance 1870 to 2010 based on twentieth century reanalysis, and links with global climate forcing. *J. Geophys. Res.* **116**, D24121 (2011).
39. Vernon, C. L. *et al.* Surface mass balance model intercomparison for the Greenland ice sheet. *Cryosphere* **7**, 599–614 (2013).
40. Janssens, I. and Huybrechts, P. The treatment of meltwater retention in mass-balance parameterizations of the Greenland ice sheet. *Ann. Glaciol.* **31**, 133–140 (2000).



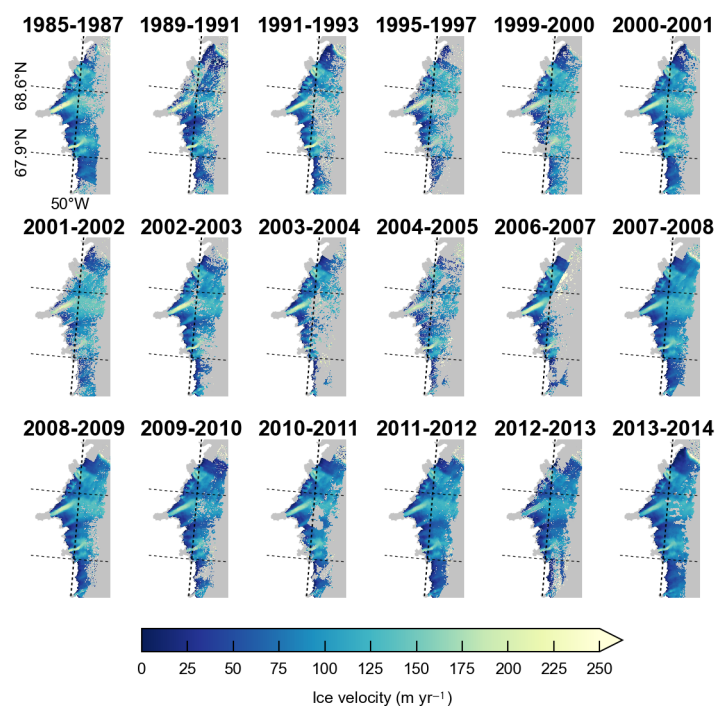
Extended Data Figure 1 | Sensitivity of extracted ice motion to variations in baseline duration. For each period are shown: **a**, the average start day-of-year (DOY) of all pairs used in the period; **b**, the average baseline duration of all pairs used in the period; **c**, the proportion of the baseline duration that is attributable to summer, which is defined as 1 May to 31 August; and **d**, the annual velocity that would be expected in the ablation zone of the Leverett glacier catchment, based on the average proportion of summer versus winter and the average baseline duration for each year.



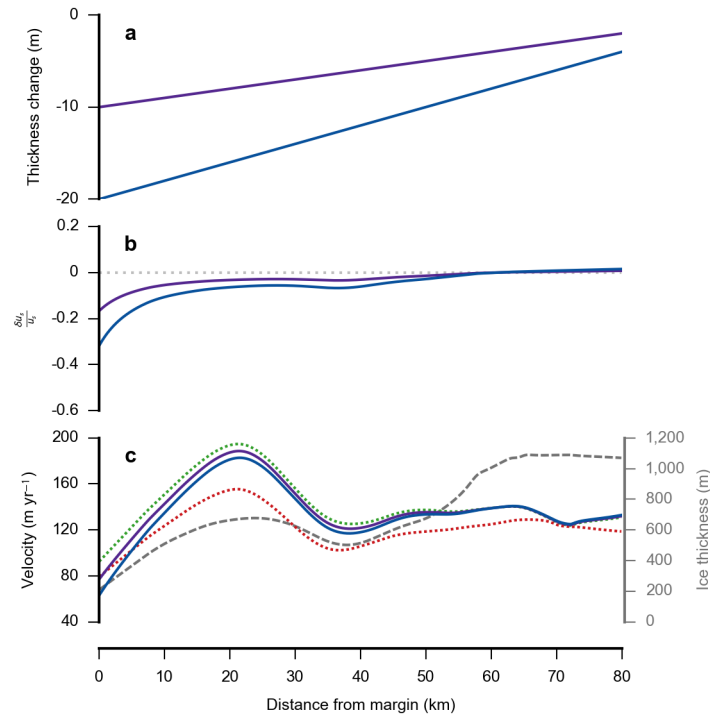
Extended Data Figure 2 | Statistical significance of three different periods of surface meltwater production. Hypothesis test of the Wilcoxon rank sum test at 95% confidence (see Methods), showing that three periods of surface melt separated by the specified dates have statistically different medians (outlined white region). The colouring shows the residuals of a three-trend linear segmented regression model fitted to melting at each possible combination of two break dates, expressed as the root-mean-squared error. The grey area is shaded as such for simplicity; it would otherwise be a mirror image of the coloured area.



Extended Data Figure 3 | Statistical significance of two different periods of ice motion. **a**, Hypothesis test of the Wilcoxon rank sum test for equal medians, testing the probability that the two populations (separated by the specified date) are similar, with 95% confidence. 0 signifies that the hypothesis of equal medians cannot be rejected, and 1 signifies that the hypothesis of equal medians can be rejected. **b**, Residuals shown as the sum-of-squares (m yr^{-1})² of a two-trend model fitted to velocities at each possible break date.



Extended Data Figure 4 | Ice velocities during each period. The velocities have uncertainties $< 60 \text{ m yr}^{-1}$ and were observed across at least 30% of the study area in each period (see Methods).



Extended Data Figure 5 | Impact of changing ice geometry on ice motion. Ice thinning of 10 m (purple) and 20 m (blue) at the ice margin, through to 0 m at 100 km inland, was applied to transect A. **a**, Prescribed change in ice thickness over transect length. **b**, The ratio of velocity change, calculated from

equation (4). **c**, Left axis, observed ice velocity during 1985–94 (dotted green) and 2007–14 (dotted red). Modelled velocities in 2014 (solid lines) for the prescribed ice thicknesses. Right axis, ice thickness (dashed grey)²⁴.

Extended Data Table 1 | Statistical relationship between melting and ice motion

Years preceding period (N)	Period and preceding N years		Preceding N years only	
	R^2	p	R^2	p
0	0.08 ⁽¹⁾	0.23	-	-
1	0.23	0.04	0.28	0.02
2	0.35	< 0.01	0.48	< 0.01
3	0.41	< 0.01	0.50	< 0.01
4	0.44	< 0.01	0.50	< 0.01

The results of linear regression analysis carried out between all periods of ice motion in Fig. 2b and different estimates of temporally coincident⁽¹⁾ and antecedent meltwater production (see Methods).

The earliest unequivocally modern humans in southern China

Wu Liu^{1*}, María Martín-Torres^{2,3,4*}, Yan-jun Cai⁵, Song Xing¹, Hao-wen Tong¹, Shu-wen Pei¹, Mark Jan Sier^{4,6,7}, Xiao-hong Wu⁸, R. Lawrence Edwards⁹, Hai Cheng¹⁰, Yi-yuan Li¹¹, Xiong-xin Yang¹², José María Bermúdez de Castro^{2,4} & Xiu-jie Wu^{1*}

The hominin record from southern Asia for the early Late Pleistocene epoch is scarce. Well-dated and well-preserved fossils older than ~45,000 years that can be unequivocally attributed to *Homo sapiens* are lacking^{1–4}. Here we present evidence from the newly excavated Fuyan Cave in Daoxian (southern China). This site has provided 47 human teeth dated to more than 80,000 years old, and with an inferred maximum age of 120,000 years. The morphological and metric assessment of this sample supports its unequivocal assignment to *H. sapiens*. The Daoxian sample is more derived than any other anatomically modern humans, resembling middle-to-late Late Pleistocene specimens and even contemporary humans. Our study shows that fully modern morphologies were present in southern China 30,000–70,000 years earlier than in the Levant and Europe^{5–7}. Our data fill a chronological and geographical gap that is relevant for understanding when *H. sapiens* first appeared in southern Asia. The Daoxian teeth also support the hypothesis that during the same period, southern China was inhabited by more derived populations than central and northern China. This evidence is important for the study of dispersal routes of modern humans. Finally, our results are relevant to exploring the reasons for the relatively late entry of *H. sapiens* into Europe. Some studies have investigated how the competition with *H. sapiens* may have caused Neanderthals' extinction (see ref. 8 and references therein). Notably, although fully modern humans were already present in southern China at least as early as ~80,000 years ago, there is no evidence that they entered Europe before ~45,000 years ago. This could indicate that *H. neanderthalensis* was indeed an additional ecological barrier for modern humans, who could only enter Europe when the demise of Neanderthals had already started.

The Fuyan Cave (25° 39' 02.7" N, 111° 28' 49.2" E; 232 m above sea level) is located in Tangbei Village, Daoxian County, Hunan Province, southern China (Fig. 1). It is part of a large multi-genesis pipeline-type karst system that contains several connected and stacked caves (Supplementary Information A), and covers an area of more than 3,000 m². The investigation and excavations were conducted at three regions in the cave, regions I, II and III (Extended Data Fig. 1). From 2011 to 2013, systematic excavations yielded 47 human teeth and an abundant fossil mammalian assemblage (Fig. 2 and Extended Data Figs 2 and 3).

Four clear stratigraphic layers were consistently identified in the whole excavated regions (regions I, II and III), with a total thickness of more than 250 cm (Fig. 1). All the hominin and mammalian fossils were found in layer 2 of region I (mammals) and region II (mammals and humans), although three human teeth (DX1, DX2 and DX6) and a

small amount of mammalian fossils were found out of context as surface findings during the first year of excavation. The stratigraphic sequence of region II, from top to bottom, is described as follows: (1) layer 1: continuous brown-grey and brown-yellow flowstone/calcite-cemented deposit with a maximum thickness of 20 cm; (2) layer 2: brown-yellow and grey fine sandy clay of 20–50 cm in thickness that contains a large amount of mammalian fossils and the hominin teeth; (3) layer 3: brown and grey sandy gravel of 80–100 cm in thickness; and (4) layer 4: grey-yellow and brown-yellow silt and clay with calcareous breccia imbedded. This layer is more than 100 cm in thickness as the bottom has not been reached yet.

At present, no stone tools have been found. The hominin and most of the faunal elements consist exclusively of teeth, and many of them present root alterations mostly due to the effects of calcium dissolution and some rodent gnawing (Supplementary Information B). The mammalian fossil assemblage from the Daoxian site is typical of Late Pleistocene in southern China, and is composed of 38 species including 5 extinct large mammals such as *Ailuropoda baconi*, *Crocota ultima*, *Stegodon orientalis*, *Megatapirus augustus* and *Sus* sp. (Extended Data Table 1 and Supplementary Information C). The radiocarbon age older than 43,000 calibrated years BP (43 kyr cal BP) obtained for one of the faunal remains (see Supplementary Information D) supports its pre-late Late Pleistocene age.

During the excavations, we collected nine samples of speleothem fragments from layers 2 to 3 (FYS-1 to FYS-9) at regions I and II, and two subsamples (FYS-S1 to FYS-S2) from a small stalagmite that grew on the top of layer 1 (Fig. 1, Extended Data Fig. 1 and Supplementary Information E). These samples were carefully preprocessed to single out the clean portion for ²³⁰Th dating, and then analysed at the Isotope Lab of University of Minnesota using the multicollector-inductively coupled plasma-mass spectrometry (MC-ICP-MS) dating technique⁹. Eight speleothem fragments from layer 2 yielded Middle to Late Pleistocene ages ranging from ~556 kyr BP to 120.7 kyr BP, and one sample collected from layer 3 (FYS-9) provided an age older than 600 kyr BP and thus, beyond the limit of the ²³⁰Th dating method (Table 1). The two subsamples from the small stalagmite give an age of 80.1 ± 1.2 kyr BP and 79.5 ± 2.8 kyr BP (mean ± 2 s.d.), respectively.

The calcitic floor (layer 1) is encrusted on layer 2, and is continuous across the excavated regions, preventing younger material from being introduced into the underlying deposits (see Supplementary Information, Cave Tour). The abundant and extensive distribution of the fauna and human teeth across the cave makes re-deposition of layer 2 highly unlikely. In addition, palaeomagnetic and rock-magnetic analysis of a sample layer 1 at region IIA confirms that

¹Key Laboratory of Vertebrate Evolution and Human Origins of Chinese Academy of Sciences, Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Sciences, Beijing 100044, China. ²UCL Anthropology, 14 Tavistock Street, London WC1H 0BW, UK. ³Departamento de Ciencias Históricas y Geografía, University of Burgos, Hospital del Rey, s/n. 09001 Burgos, Spain. ⁴Centro Nacional de Investigación sobre la Evolución Humana (CENIEH), Paseo Sierra de Atapuerca 3, 09002 Burgos, Spain. ⁵State Key Laboratory of Loess and Quaternary Geology, Institute of Earth Environment, Chinese Academy of Sciences, Xian 710075, China. ⁶Paleomagnetic Laboratory 'Fort Hoofddijk', Department of Earth Sciences, Faculty of Geosciences, Utrecht University, Budapestlaan 17, 3584 CD Utrecht, The Netherlands. ⁷Faculty of Archaeology, Leiden University, PO Box 9515, 2300 RA Leiden, The Netherlands. ⁸School of Archaeology and Museology, Peking University, Beijing 100871, China. ⁹Department of Geology and Geophysics, University of Minnesota, Minneapolis, Minnesota 55455, USA. ¹⁰Institute of Global Environmental Change, Xi'an Jiaotong University, Xi'an 710049, China. ¹¹Institute of Cultural Relics and Archaeology, Hunan Province, Changsha 410008, China. ¹²Cultural Relics Administration of Daoxian County, Daoxian 425300, China.

*These authors contributed equally to this work.

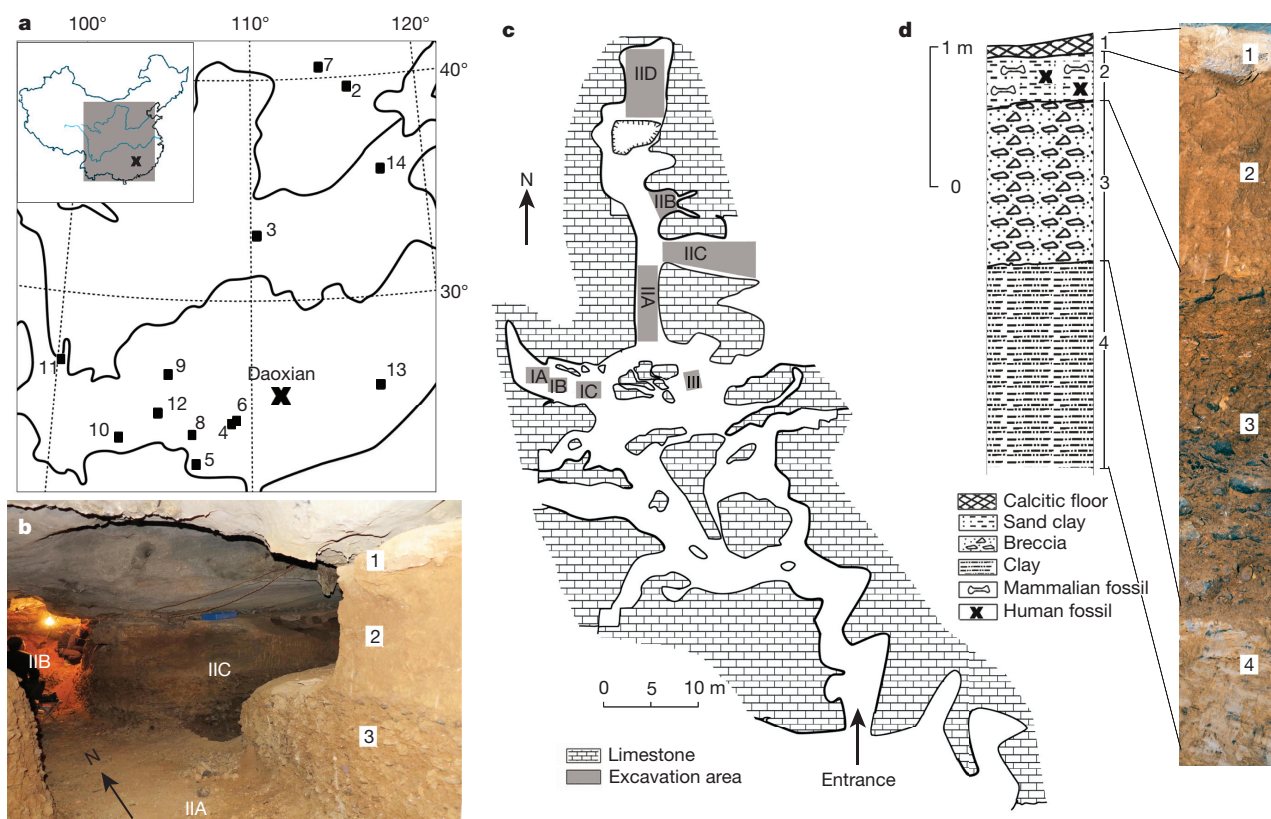


Figure 1 | Geographical location and stratigraphy of the Daoxian site.

a, Location of the Daoxian site. Late Middle Pleistocene and Late Pleistocene localities with human remains that have been included in the morphological and/or metric comparison with Daoxian are also marked on the map. 2: Tianyuan Cave; 3: Huanglong Cave; 4: Liujiang; 5: Zhiren Cave; 6: Tubo; 7: Xujiayao; 8: Luna; 9: Chuandong; 10: Malu Cave; 11: Lijiang; 12: Longlin; 13: Huli Cave; and 14: Xintai. The map is adapted from the original Chinese map

from National Administration of Surveying, Mapping and Geoinformation of China (<http://219.238.166.215/mcp/index.asp>). **b**, General view of the interior of the cave and the spatial relationship of regions IIA, IIB and IIC, with some of the layers marked. **c**, Plan view of the excavation area. **d**, Detail of the stratigraphic layers of region II of the Daoxian site. All human fossils come from layer 2.

the flowstone that caps the fossil-bearing layer 2 remains *in situ* (Supplementary Information F and Extended Data Fig. 4). Therefore, the dated stalagmite was formed after the fossils were buried and it provides a minimum age constraint (~ 80 kyr) for the fossils below. Because the associated fauna is typical of the Late Pleistocene, we conservatively assume that fossils are not older than ~ 120 kyr, and

the presence of hominins at Daoxian can be bracketed between 80 kyr and 120 kyr.

Daoxian teeth were compared to large dental samples of Late Pleistocene hominin fossils from Europe, Africa and Asia (Extended Data Tables 2 and 3 and Supplementary Information G and H). The Daoxian teeth are small and they consistently fall within *H. sapiens*

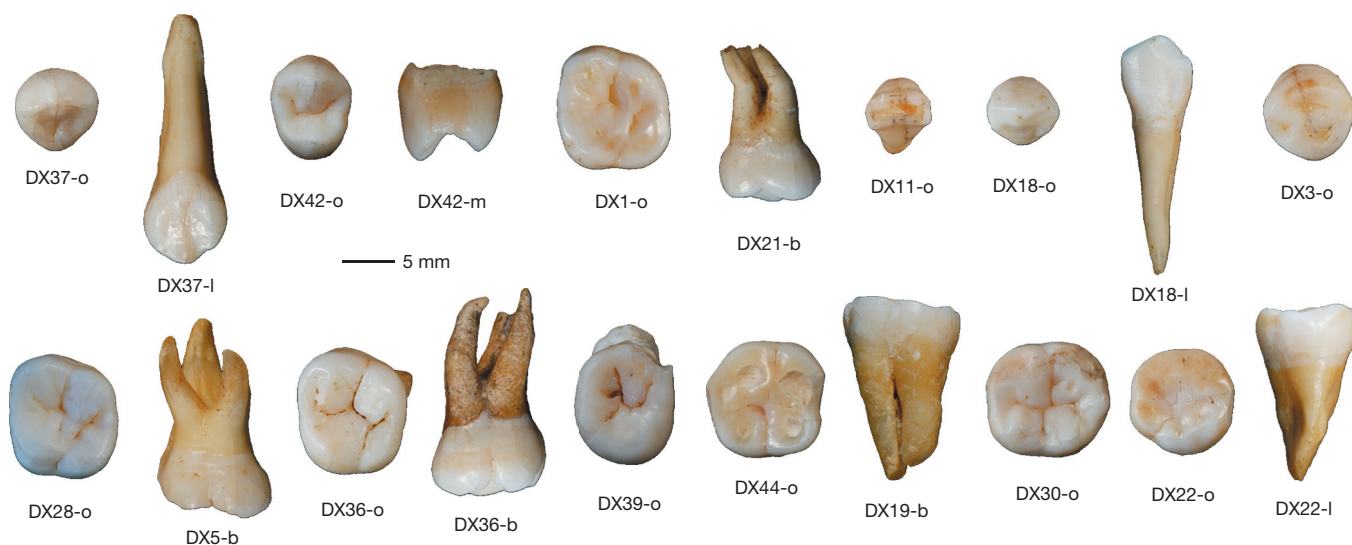


Figure 2 | Daoxian human teeth (selection). See Extended Data Table 2 for detailed information about each tooth. b, buccal; d, distal; l, lingual; m, mesial; o, occlusal. Credits: S.X. and X.-J.W.

Table 1 | The ^{230}Th ages of the Daoxian site

Sample ID	Region/layer	^{238}U (ppb)	^{232}Th (ppt)	$^{230}\text{Th}/^{232}\text{Th}$ (atomic $\times 10^{-6}$)	$\delta^{234}\text{U}^*$ (measured)	$^{230}\text{Th}/^{238}\text{U}$ (activity)	^{230}Th age (kyr BP) (uncorrected)	$\delta^{234}\text{U}_{\text{initial}}^\dagger$ (corrected)	^{230}Th age (kyr BP) ‡ (corrected)
FYS-S1	IID/layer 1	133.3 \pm 0.3	9,117 \pm 183	176.5 \pm 3.6	353.1 \pm 4.5	0.7326 \pm 0.0033	81.5 \pm 0.7	443 \pm 6	80.1 \pm 1.2
FYS-S2	IID/layer 1	285.9 \pm 0.4	55,026 \pm 1,102	64.0 \pm 1.3	356.2 \pm 3.2	0.7467 \pm 0.0026	83.4 \pm 0.5	446 \pm 5	79.5 \pm 2.8
FYS-1	IIA/layer 2	428.2 \pm 0.7	98,699 \pm 1,976	59.4 \pm 1.2	54.3 \pm 2.1	0.8302 \pm 0.0017	164.7 \pm 1.2	85 \pm 3	158.3 \pm 4.6
FYS-2	IIA/layer 2	10,747.9 \pm 69.1	27,564 \pm 552	7,463.3 \pm 149.9	633.0 \pm 4.2	1.1609 \pm 0.0077	121.0 \pm 1.5	891 \pm 7	121.0 \pm 1.5
FYS-3	IC/layer 2	126.0 \pm 0.2	8,675 \pm 174	263.5 \pm 5.3	75.6 \pm 2.5	1.1000 \pm 0.0027	558.3 \pm 62.8	364 \pm 67	556.8 \pm 61.9
FYS-4	IB/layer 2	1,608.5 \pm 5.5	889 \pm 18	29,237 \pm 588	401.0 \pm 3.5	0.9800 \pm 0.0035	120.7 \pm 0.9	564 \pm 5	120.7 \pm 0.9
FYS-5	IB/layer 2	260.0 \pm 0.4	41,356 \pm 828	122.7 \pm 2.5	173.3 \pm 2.7	1.1837 \pm 0.0026	351.5 \pm 8.1	463 \pm 13	348.3 \pm 8.2
FYS-6	IA/layer 2	120.6 \pm 0.2	81,358 \pm 1,629	22.6 \pm 0.5	171.4 \pm 2.7	0.9236 \pm 0.0027	158.5 \pm 1.4	256 \pm 10	141.8 \pm 12.1
FYS-7	IID/layer 2	87.4 \pm 0.1	12,302 \pm 246	120.3 \pm 2.5	187.8 \pm 5.1	1.0276 \pm 0.0055	196.1 \pm 3.8	324 \pm 10	192.9 \pm 4.3
FYS-8	IID/layer 2	78.4 \pm 0.2	20,571 \pm 413	55.2 \pm 1.1	157.4 \pm 7.7	0.8786 \pm 0.0044	147.1 \pm 2.7	234 \pm 12	140.7 \pm 5.2
FYS-9	IIC/layer 3	267.6 \pm 0.4	20,907 \pm 419	10,618.9 \pm 226.1	147.6 \pm 3.2	2.5165 \pm 0.3707	>600	–	–

* $\delta^{234}\text{U} = ((^{234}\text{U}/^{238}\text{U})_{\text{activity}} - 1) \times 1,000$.

$^\dagger \delta^{234}\text{U}_{\text{initial}}$ was calculated based on ^{230}Th age (T), that is, $\delta^{234}\text{U}_{\text{initial}} = \delta^{234}\text{U}_{\text{measured}} \times \exp(\lambda_{234} \times T)$.

‡ Corrected ^{230}Th ages assume the initial $^{230}\text{Th}/^{232}\text{Th}$ atomic ratio of $4.4 \pm 2.2 \times 10^{-6}$. Those are the values for a material at secular equilibrium, with the bulk earth $^{232}\text{Th}/^{238}\text{U}$ value of 3.8. The errors are arbitrarily assumed to be 50%.

BP stands for 'before present', in which 'present' is defined as the year 1950 AD. Values are mean \pm 2 s.d.

variability (Fig. 3 and Extended Data Fig. 5). They are generally smaller than other Late Pleistocene specimens from Africa and Asia, and closer to European Late Pleistocene samples and contemporary modern humans. Both the crown and the root of Daoxian teeth show typical morphologies for *H. sapiens* (Fig. 2 and Extended Data Fig. 6), with simplified occlusal and labial/buccal surfaces and short and slender roots. The presence of moderate basal bulging as well as longitudinal grooves in the buccal surface of canines, premolars and molars from other Late Pleistocene samples such as Xujiayao, Huanglong Cave, Qafzeh or Dolni Vestonice make Daoxian teeth morphologically closer to middle-to-late Late Pleistocene and even contemporary human samples (Extended Data Fig. 6). Canine and molar roots are gracile and barely divergent, differing from the stout and robust root systems of Tubo or Xujiayao localities¹⁰ where radicals do not narrow towards the tip (Extended Data Fig. 6). Indeed, the convergent apices of the molar buccal roots appear as a typical feature in contemporary *H. sapiens*. M^1 molars are also typical of *H. sapiens* and unlike the rhomboidal contour displayed by *H. neanderthalensis*¹¹ or the buccolingually elongated shape of Asian *H. erectus*^{12–14}. The relative cusp and occlusal polygon areas of the Daoxian M^1 molars follow the *H. sapiens*

pattern and they only differ by 0.6% to 1.1% from modern Chinese populations (Extended Data Table 4). Interestingly, Qafzeh M^1 s are comparatively less derived than Daoxian, showing a departure from the typical *H. sapiens* pattern of cusp proportions and angles that were previously noticed¹¹. The occlusal morphology of Daoxian M^2 and M^3 s is also simple, and both the metacone and the hypocone are strongly reduced as it is typical of *H. sapiens*. The lack of labial convexity, shovel shape and tuberculum dentale, as well as the gracile root of the Daoxian incisor I_2 resemble that of contemporary and Late Pleistocene *H. sapiens* and differs from Neanderthals. However, Dolni Vestonice specimens display higher labial convexity, and Qafzeh and Huanglong Cave I_2 s present a more prominent basal eminence, making Daoxian I_2 closer to contemporary humans rather than other Late Pleistocene samples. The two Daoxian P_3 premolars show a slightly asymmetric oval contour due to the disto-lingual projection of a small platform-like talonid without accessory cusps. Overall, the crown morphology together with the expression of a slender single root of Daoxian P_3 s is closer to *H. sapiens* and differs from the typical Neanderthal conformation, with compressed and centred occlusal polygon and lingually displaced metaconid¹⁵. Lower molars lack the typically

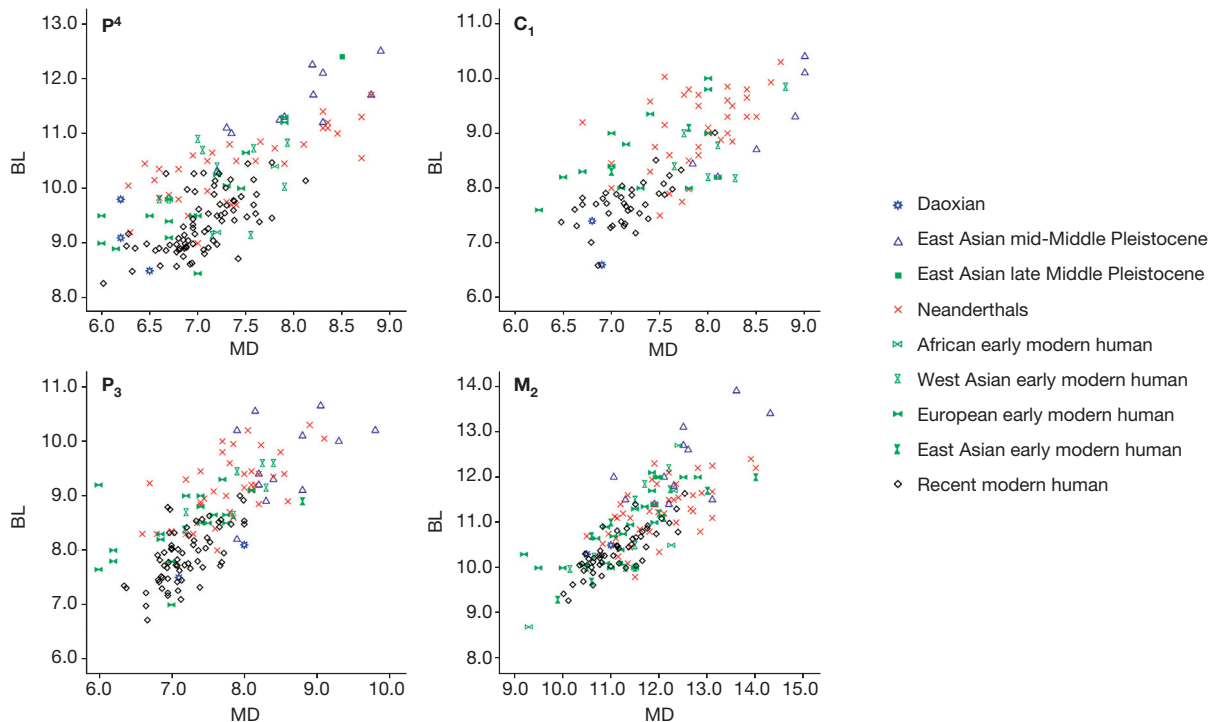


Figure 3 | Metric comparison of Daoxian teeth. Bivariate plots of the mesiodistal (MD) and buccolingual (BL) dimensions of the Daoxian upper fourth premolar (P^4), lower canine (C_1), lower third premolar (P_3) and lower second molar (M_2).

Neanderthal combination of a pit-like anterior fovea with a continuous mid-trigonid crest^{16,17}. This, together with the reduction of the hypoconulid and the expression of an X-pattern in all the M₂ and M₃ where this feature could be recorded make Daoxian lower molars morphologically closer to anatomically and contemporary modern humans¹⁸. In addition, no signs of taurodontism are present. Finally, the occlusal morphology of the two upper second deciduous molars (dm²s) from Daoxian is simple, and the occlusal outline relatively squared and unlike the typical skewed contour of Neanderthals. Roots are thin and diverge as is typical in deciduous teeth, and similar to the patterns usually found in fossil and contemporary *H. sapiens*. Thus, the morphological and metric comparison of the Daoxian dental sample allows its unequivocal attribution to *H. sapiens*, and they present particular resemblances to late Late Pleistocene samples and contemporary modern humans.

At present, the earliest unambiguous evidence of *H. sapiens* fossils eastward of the Arabian Peninsula comes from Tianyuan Cave, in northern China¹⁹, Niah Cave in Borneo⁴ and Lake Mungo in Australia²⁰ dated to ~40,000–50,000 years. The retention of primitive features in Qafzeh and Skhul has been interpreted by many as evidence of a ‘failed’ dispersal^{21,22}, and several studies have recently suggested that an earlier and southern route may have been indeed more favourable for hominin expansion^{2,23–25}. However, these and other related hypotheses were lacking the support of clear evidence of modern human occupation outside Africa (excluding the Levant) during the early Late Pleistocene. The fragmentary nature and/or the mosaic of modern and archaic features of remains such as those from the Zhiren Cave have prevented a unanimous acceptance of its taxonomic status^{1,26}. This, together with the contested chronological-stratigraphic frame of some of the Asian hominin findings (see ref. 2 for a review), make the Daoxian teeth the earliest and soundest evidence of definitely modern humans in southern China at least 80 kyr ago. The Daoxian evidence may finally change the scepticism that most hypotheses considering the presence of *H. sapiens* in the early Late Pleistocene in China have been subjected to.

While the Daoxian findings would support the presence of fully modern populations in southern China during the early Late Pleistocene, the Xujiayao¹⁰ and Denisova evidence²⁷ points to considerably more primitive hominins in the northern latitudes. Similarly, the dental morphology of the late Middle Pleistocene hominin from Panxian Dadong in southern China already exhibits some derived features²⁸ that are absent in other roughly contemporaneous Asian populations of higher latitudes, such as those from Zhoukoudian, Hexian or Chaoxian¹⁰. This evidence could support different origins and/or dispersal routes for modern humans across Asia^{23,24}.

Finally, while fully modern humans succeeded to disperse throughout Asia during the early Late Pleistocene, they failed to do so in Europe until 35,000–75,000 years later. Thus, we should not rule out the possibility that *H. neanderthalensis* was for a long time an additional barrier for modern humans’ expansion, who could only settle in Europe when Neanderthal populations started to fade.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 9 May; accepted 9 September 2015.

Published online 14 October 2015.

- Dennell, R. Early *Homo sapiens* in China. *Nature* **468**, 512–513 (2010).
- Dennell, R. in *Southern Asia, Australia and the Search for Human Origins* (eds Dennell, R. & Porr, M.) 33–50 (Cambridge Univ. Press, 2014).
- Storm, P. *et al.* U-series and radiocarbon analyses of human and faunal remains from Wajak, Indonesia. *J. Hum. Evol.* **64**, 356–365 (2013).
- Barker, G. *et al.* The ‘human revolution’ in lowland tropical Southeast Asia: the antiquity and behavior of anatomically modern humans at Niah Cave (Sarawak, Borneo). *J. Hum. Evol.* **52**, 243–261 (2007).
- Hershkovitz, I. *et al.* Levantine cranium from Manot Cave (Israel) foreshadows the first European modern humans. *Nature* **520**, 216–219 (2015).

- Grine, F. E. *et al.* Late Pleistocene human skull from Hofmeyr, South Africa, and modern human origins. *Science* **315**, 226–229 (2007).
- Benazzi, S. *et al.* The makers of the Protoaurignacian and implications for Neanderthal extinction. *Science* **348**, 793–796 (2015).
- Villa, P. & Roebroeks, W. Neanderthal demise: an archaeological analysis of the modern human superiority complex. *PLoS ONE* **9**, e96424 (2014).
- Cheng, H. *et al.* Improvements in ²³⁰Th dating, ²³⁰Th and ²³⁴U half-life values, and U–Th isotopic measurements by multi-collector inductively coupled plasma mass spectrometry. *Earth Planet. Sci. Lett.* **371–372**, 82–91 (2013).
- Xing, S., Martín-Torres, M., Bermúdez de Castro, J. M., Wu, X. & Liu, W. Hominin teeth from the early Late Pleistocene site of Xujiayao, Northern China. *Am. J. Phys. Anthropol.* **156**, 224–240 (2015).
- Bailey, S. E. A morphometric analysis of maxillary molar crowns of Middle-Late Pleistocene hominins. *J. Hum. Evol.* **47**, 183–198 (2004).
- Xing, S. *et al.* Middle Pleistocene hominin teeth from Longtan Cave, Hexian, China. *PLoS ONE* **9**, 3114265 (2014).
- Kaifu, Y. Advanced dental reduction in Javanese *Homo erectus*. *Anthropol. Sci.* **114**, 35–43 (2006).
- Kaifu, Y. *et al.* Taxonomic affinities and evolutionary history of the early Pleistocene hominids of Java: dentognathic evidence. *Am. J. Phys. Anthropol.* **128**, 709–726 (2005).
- Gómez-Robles, A. *et al.* Geometric morphometric analysis of the crown morphology of the lower first premolar of hominins, with special attention to Pleistocene *Homo*. *J. Hum. Evol.* **55**, 627–638 (2008).
- Bailey, S. E. *Neanderthal Dental Morphology: Implications for Modern Human Origins*. PhD thesis, Arizona State Univ. (2002).
- Martín-Torres, M., Bermúdez de Castro, J. M., Gómez-Robles, A., Prado-Simón, L. & Arsuaga, J. L. Morphological description and comparison of the dental remains from Atapuerca-Sima de los Huesos site (Spain). *J. Hum. Evol.* **62**, 7–58 (2012).
- Martín-Torres, M. *et al.* Dental evidence on the hominin dispersals during the Pleistocene. *Proc. Natl Acad. Sci. USA* **104**, 13279–13282 (2007).
- Shang, H. & Trinkaus, E. *The Early Modern Human from Tianyuan Cave, China* (Texas A&M Univ. Press, 2010).
- Bowler, J. M. *et al.* New ages for human occupation and climatic change at Lake Mungo, Australia. *Nature* **421**, 837–840 (2003).
- Mellars, P. Why did modern human populations disperse from Africa ca. 60,000 years ago? A new model. *Proc. Natl Acad. Sci. USA* **103**, 9381–9386 (2006).
- Oppenheimer, S. The great arc of dispersal of modern humans: Africa to Australia. *Quat. Int.* **202**, 2–13 (2009).
- Armitage, S. J. *et al.* The Southern Route “Out of Africa”: evidence for an early expansion of modern humans into Arabia. *Science* **331**, 453–456 (2011).
- Reyes-Centeno, H. *et al.* Genomic and cranial phenotype data support multiple modern human dispersals from Africa and a southern route into Asia. *Proc. Natl Acad. Sci. USA* **111**, 7248–7253 (2014).
- Westaway, K. E. *et al.* Age and biostratigraphic significance of the Punung Rainforest Fauna, East Java, Indonesia, and implications for *Pongo* and *Homo*. *J. Hum. Evol.* **53**, 709–717 (2007).
- Liu, W. *et al.* Human remains from Zhirendong, South China, and modern human emergence in East Asia. *Proc. Natl Acad. Sci. USA* **107**, 19201–19206 (2010).
- Reich, D. *et al.* Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**, 1053–1060 (2010).
- Liu, W. *et al.* Late Middle Pleistocene hominin teeth from Panxian Dadong, South China. *J. Hum. Evol.* **64**, 337–355 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work has been supported by the grants from the Chinese Academy of Sciences (KZD-EW-03, XDA05130101, GJHZ201314), National Natural Science Foundation of China (41272034, 41302016, 41271229), Netherlands Organisation for Scientific Research (NWO-ALW 823.01.003), Dirección General de Investigación of the Spanish Ministerio de Educación y Ciencia (CGL2012-38434-C03-02, and Acción Integrada España Francia HF2007-0115), Consejería de Educación de Junta de Castilla y León (CEN074A12-2) and The Leakey Foundation (through the support of G. Getty and D. Crook). We are grateful to several people who have provided access to comparative materials and/or advice in several aspects of the manuscript: R. Blasco, J. Rosell, J. M. Parés, M. Salesa, A. Tarriño, C. Saiz, I. Hershkovitz, A. Viallet, M. A. de Lumley, C. Bernis, J. Rascón and J. Svoboda. We are also grateful to Y.-S. Lou, L.-M. Zhang and P.-P. Wei who participated in the excavations at the Daoxian site.

Author Contributions X.-J.W., W.L. and M.M.-T. are the corresponding authors and have contributed equally to this work. X.-J.W. and W.L. are directing the Daoxian research project. W.L., M.M.-T., S.X., X.-J.W. and J.M.B.d.C. performed the anthropological study of the Daoxian human teeth. Y.-J.C. and S.-W.P. conducted the geological studies of the Daoxian site. Y.-J.C., R.L.E. and H.C. conducted the U–Th dating of the speleothem and stalagmite samples collected from the cave. M.J.S. conducted the palaeomagnetic analysis. X.-H.W. conducted the radiocarbon dating. H.-W.T. conducted the study of the faunal remains. X.-J.W., X.-X.Y., Y.-Y.L., W.L., Y.-J.C., H.-W.T. and S.-W.P. participated in the field research.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.M.-T. (maria.martinon-torres@ucl.ac.uk), W.L. (liuwu@ivpp.ac.cn) or X.-J.W. (wuxijie@ivpp.ac.cn).

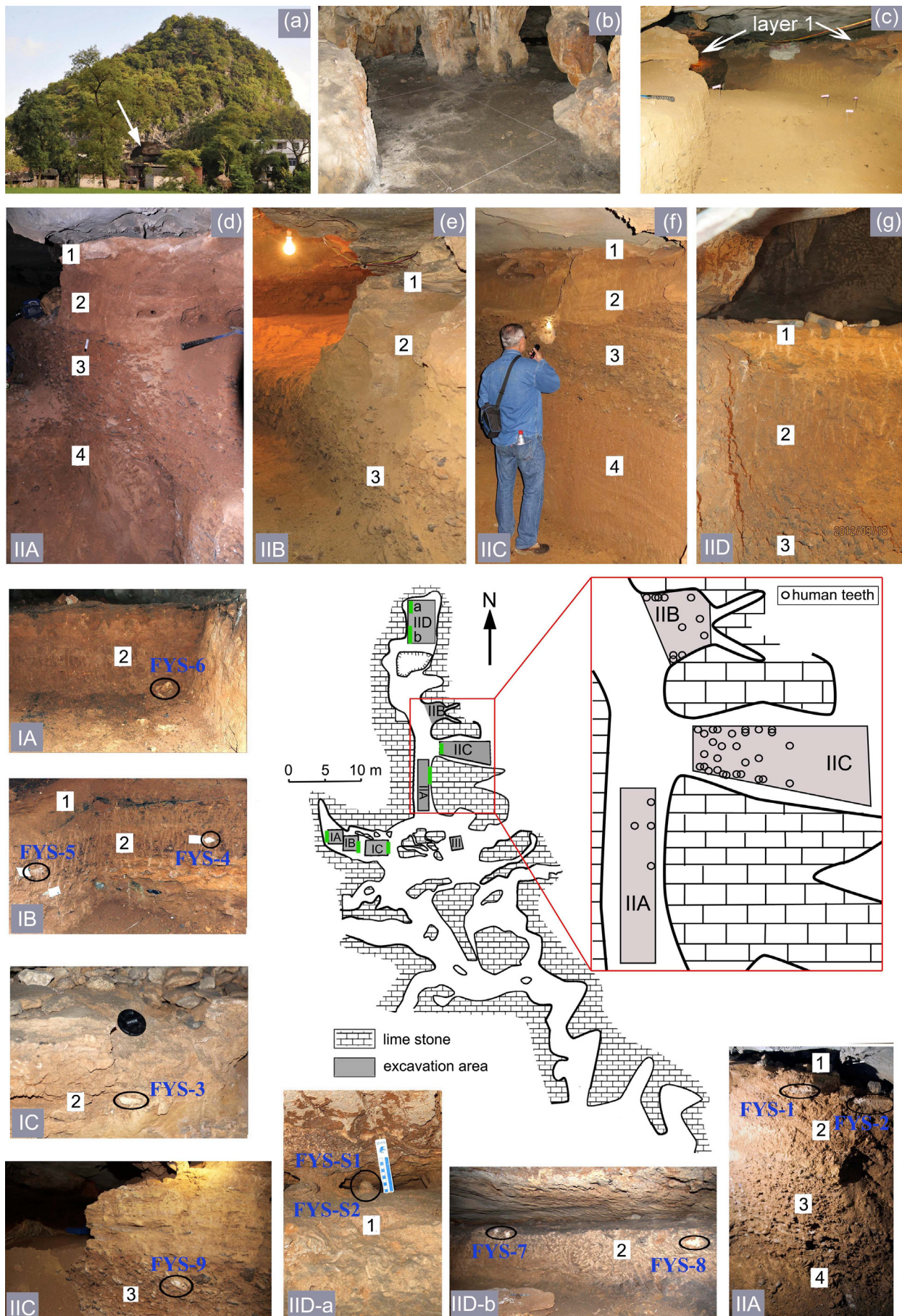
METHODS

Description of the morphological features of the Daoxian hominin teeth follow the terminology usually used in dental studies^{17,29,30}. To assess the morphological affinities of Daoxian teeth, we compared them to other Late Pleistocene samples from Africa, Asia and Europe (including Neanderthals), as well as a large contemporary *H. sapiens* sample (see Extended Data Table 3). Apart from both the descriptive comparative anatomy and the mesiodistal and buccolingual comparison, in the case of the M¹ we also calculated the relative cusp and occlusal polygon size.

For the metric comparison, the crown mesiodistal and buccolingual dimensions of the Daoxian teeth were measured with a standard sliding caliper and recorded to the nearest 0.1 mm. Bivariate plots of the mesiodistal and buccolingual diameters will be provided for the metric comparison of Daoxian with other hominin samples. To explore the Daoxian hominins in the context of the Middle to Late Pleistocene evolutionary changes in China, some Middle Pleistocene hominins from China were also included.

In addition, the total crown area and relative cusp area, and relative polygon areas for upper first molars were also measured and compared to a modern Chinese population. These features are considered to be taxonomically discriminative, particularly between *H. sapiens* and *H. neanderthalensis*¹¹. The protocols for the measurement and calculation of the relative cusp areas of M¹ can be found in ref. 11.

29. Scott, G. R. & Turner, C. G. *The Anthropology of Modern Human Teeth: Dental Morphology and its Variation in Recent Human Populations* (Cambridge Univ. Press, 1997).
30. Weidenreich, F. *The Dentition of Sinanthropus Pekinensis: A Comparative Odontology of the Hominids* 1st edn (Geological Survey of China, 1937).
31. Li, Y. *et al.* A preliminary report on the 2011 excavation at Houbeishan Fuyan Cave, Daoxian, Hunan Province. *Acta Anthropol. Sinica* **32**, 133–143 (2013).
32. Wu, X. Z. *Yunxi Man: Excavation Report of the Huanglong Cave* (Science Press, 2006).
33. Jin, C. Z. *et al.* The *Homo sapiens* cave hominin site of Mulan Mountain, Jiangzhou District, Chongzhou, Guianxi with emphasis on its age. *Chin. Sci. Bull.* **54**, 3848–3856 (2009).
34. Yi, G. Liujian Man. *Fossils* **2**, 15 (1982).
35. Woo, J. Human fossils found in Kiukiang, Kwangsi, China. *Paleovertebrata Paleanthropologica* **1**, 97–104 (1959).
36. Molnar, S. Human tooth wear, tooth function and cultural variability. *Am. J. Phys. Anthropol.* **34**, 175–189 (1971).
37. White, T. D. *et al.* Pleistocene *Homo sapiens* from Middle Awash, Ethiopia. *Nature* **423**, 742–747 (2003).
38. Rightmire, G. P. & Deacon, H. J. New human teeth from Middle Stone Age deposits at Klasies River, South Africa. *J. Hum. Evol.* **41**, 535–544 (2001).
39. Bräuer, G. & Mehlman, M. J. Hominid molars from a Middle Stone Age level at the Mumba Rock Shelter, Tanzania. *Am. J. Phys. Anthropol.* **75**, 69–76 (1988).
40. Chia, L. Note on the human and some other mammalian remains from Chgyang, Hupei. *Vertebr. Palasiat.* **1**, 247–257 (1957).
41. He, J. Preliminary study on the teeth of Jinniushan archaic *Homo sapiens*. *Acta Anthropol. Sinica* **19**, 216–224 (2000).
42. Wu, M. New discoveries of human fossils in Tongzi, Guizhou. *Acta Anthropol. Sinica* **3**, 195–201 (1984).
43. Wu, M., Wang, L., Zhang, Y. & Zhang, S. Fossil human teeth and associated cultural relics from Tongzi, Guizhou Province. *Vertebr. Palasiat.* **13**, 14–23 (1975).
44. Gu, Y. Zhoukoudian New Cave Man and their living environment. *Selected Papers in Paleoanthropology* 158–171 (Science Press, 1978).
45. Bae, C. J. *et al.* Modern human teeth from Late Pleistocene Luna Cave (Guangxi, China). *Quat. Int.* **354**, 169–183 (2014).
46. Chen, D. & Qi, G. Fossil human and associated mammalian fauna found from Xizhou, Yunnan. *Vertebr. Palasiat.* **16**, 33–46 (1978).
47. Dong, X. & Fan, X. Note on human fossil teeth from Fox Cave at Qingliu. *Acta Anthropol. Sinica* **15**, 315–319 (1996).
48. Huang, S. & Zheng, L. The upper Pleistocene human tooth and mammalian fossil from Changwu, Xhaanxi. *Acta Anthropol. Sinica* **1**, 14–17 (1982).
49. Li, Y., Wu, M., Peng, S. & Zhou, S. Human tooth fossils and some mammalian remains from Tubo, Liujiang, Guangxi. *Acta Anthropol. Sinica* **3**, 322–329 (1984).
50. Li, Y., Wu, M., Peng, S. & Zhou, S. Preliminary report on the investigation of Dingmo Cave in Tiandong County, Guangxi. *Acta Anthropol. Sinica* **4**, 127–131 (1985).
51. Peng, S. & Wang, W. Fossil of human beings and mammal discovered in Longdong Cave at Longlin, Guangxi. *Ethnoarchaeol. South China* **3**, 187–292 (1990).
52. Wang, W., Huang, Q. & Zhou, S. New found human tooth fossils in Tubo, Guangxi. *Longgupo Prehistory Culture* **1**, 104–108 (1999).
53. Wang, W. & Mo, J. Human fossil teeth newly discovered in Nanshan Cave of Fusui, Guangxi. *Acta Anthropol. Sinica* **23**, 130–136 (2004).
54. Wang, W., Huang, C., Xie, S. & Yan, C. Late Pleistocene hominin teeth from the Jimuyan Cave, Pingle County, Guangxi, South China. *Quat. Sci.* **31**, 699–704 (2011).
55. Wu, X. Z., Zhao, Z., Yuan, Z. & Shen, J. Report on paleoanthropological expedition of the Northeastern part of Kwangsi. *Vertebr. Palasiat.* **6**, 408–413 (1962).
56. Wu, X. & Zong, G. A human tooth and mammalian fossils of Late Pleistocene in Wuzhuta, Xintai, Shantong. *Vertebr. Palasiat.* **11**, 105–106 (1973).
57. Yu, J. Fossil man and cultural artifacts from Chuandong, Puding County, Guizhou Province. *J. Nanjing Univ. Nat. Sci.* **20**, 145–155 (1984).
58. You, Y., Dong, X., Chen, C. & Fan, X. A fossil human tooth from Qingliu, Fujian. *Acta Anthropol. Sinica* **8**, 197–202 (1989).
59. Zheng, L. A fossil human tooth from Zhaotong, Yunnan. *Acta Anthropol. Sinica* **4**, 105–108 (1985).
60. Zhou, G. & Yi, G. On the remains from Liuzhou region, Guangxi. *Mem. Beijing Nat. His. Mus.* **20**, 1–21 (1983).
61. Sládek, V., Trinkaus, E., Hillson, S. W. & Holiday, T. W. *The People of the Pavlovian: Skeletal Catalogue and Osteometrics of the Gravettian Fossil Hominids from Dolní Věstonice and Pavlov* (Dolní Věstonice Studies, 2000).
62. Bailey, S., Glantz, M., Weaver, T. D. & Viola, B. The affinity of the dental remains from Obi-Rakhmat Grotto, Uzbekistan. *J. Hum. Evol.* **55**, 238–248 (2008).
63. Quam, R., Bailey, S. E. & Wood, B. A. Evolution of M¹ crown size and cusp proportions in the genus *Homo*. *J. Anat.* **214**, 655–670 (2009).

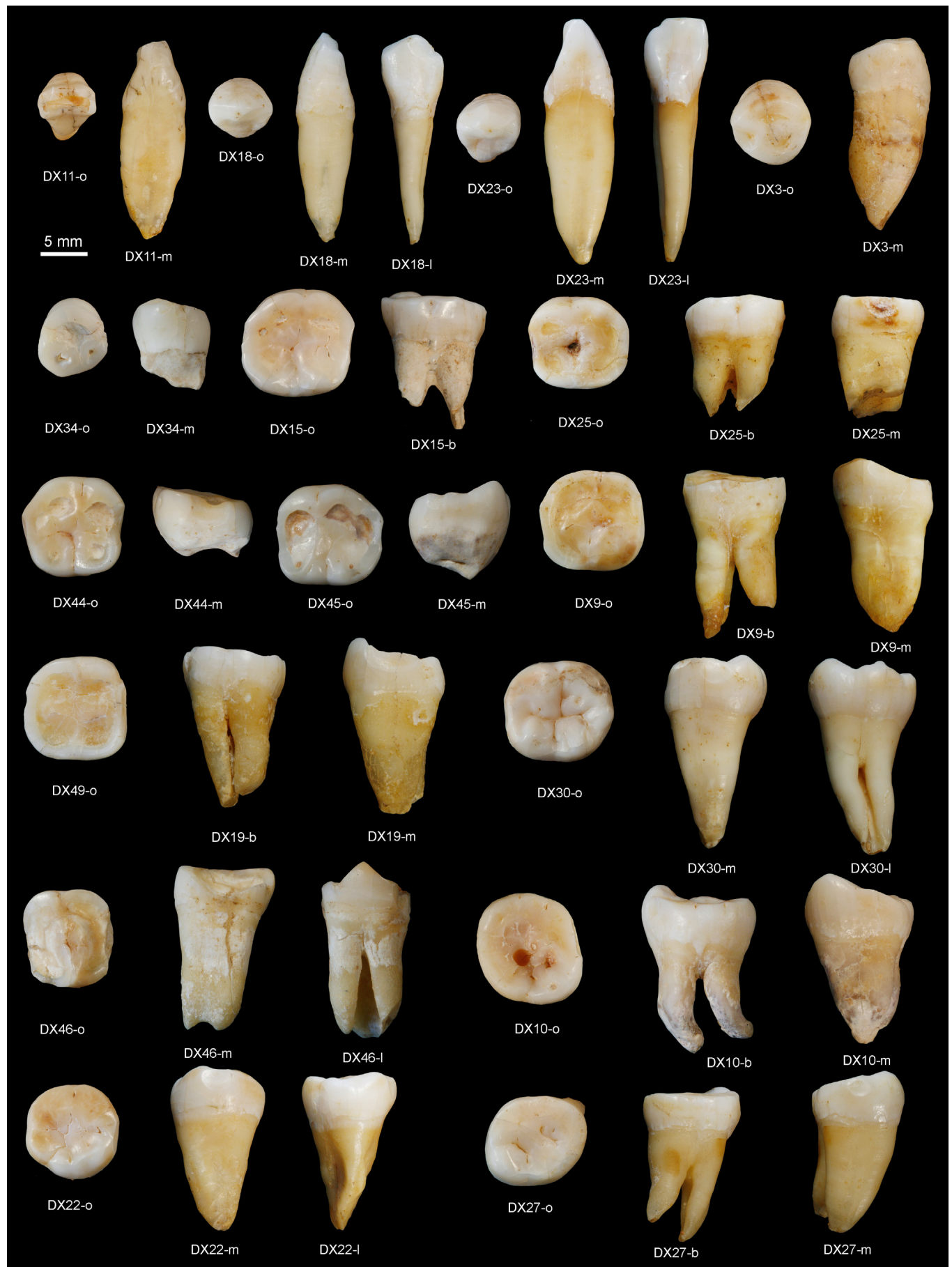


Extended Data Figure 1 | The Daoxian site. **a**, Entrance to the Fuyan (Daoxian) Cave. **b**, Image of the intact flowstone in an unexcavated area. **c**, Detail of the excavation at region IIC. Pink flags point to *in situ* human findings. **d–g**, Detail of the stratigraphy of region IIA (**d**), IIB (**e**), IIC (**f**) and IID (**g**). In the centre, plan view of the excavation area at the Daoxian Cave. The

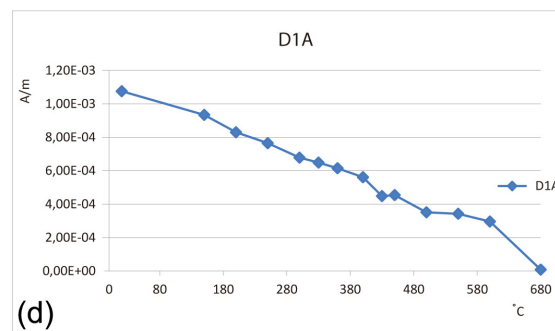
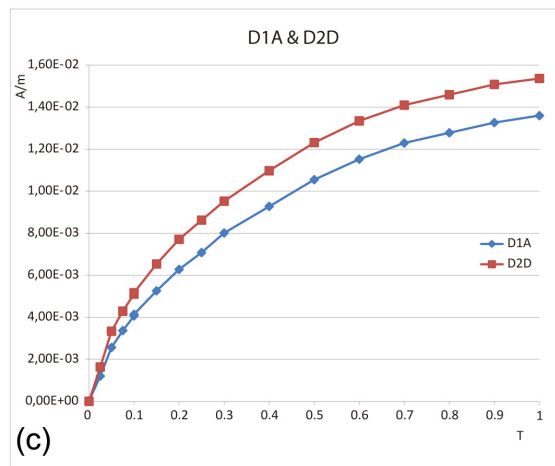
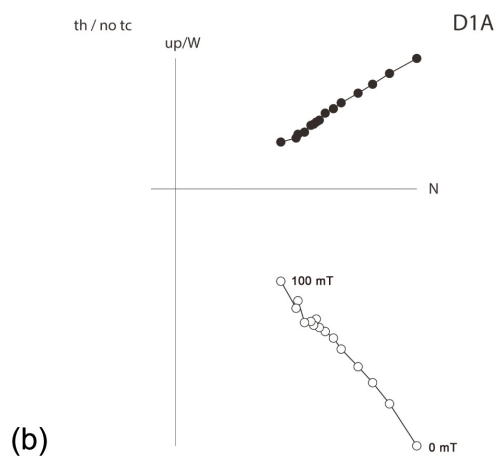
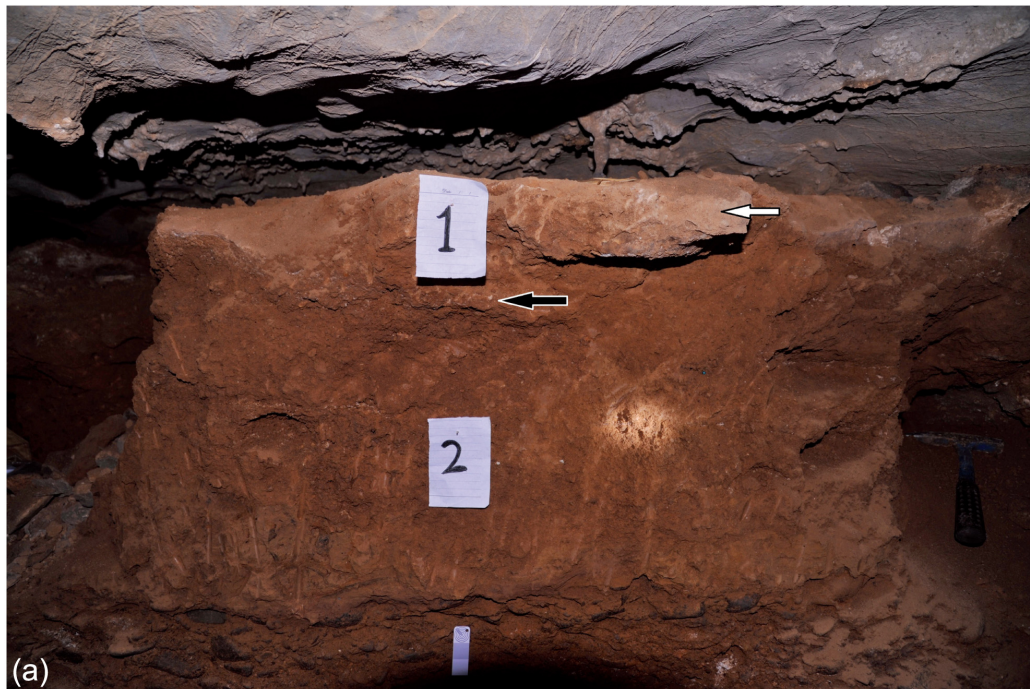
enlarged area shows the individual location of each human tooth. Lower pictures provide a detail of the location of each dating sample. FYS, speleothem fragment samples; FYS-S, stalagmite samples. For more details on the U-series results, see Table 1 and Supplementary Information E.



Extended Data Figure 2 | Daoxian upper teeth. Please see Extended Data Table 2 for detailed information. b, buccal; d, distal; l, lingual; m, mesial; o, occlusal.



Extended Data Figure 3 | Daoxian lower teeth. Please see Extended Data Table 2 for detailed information.

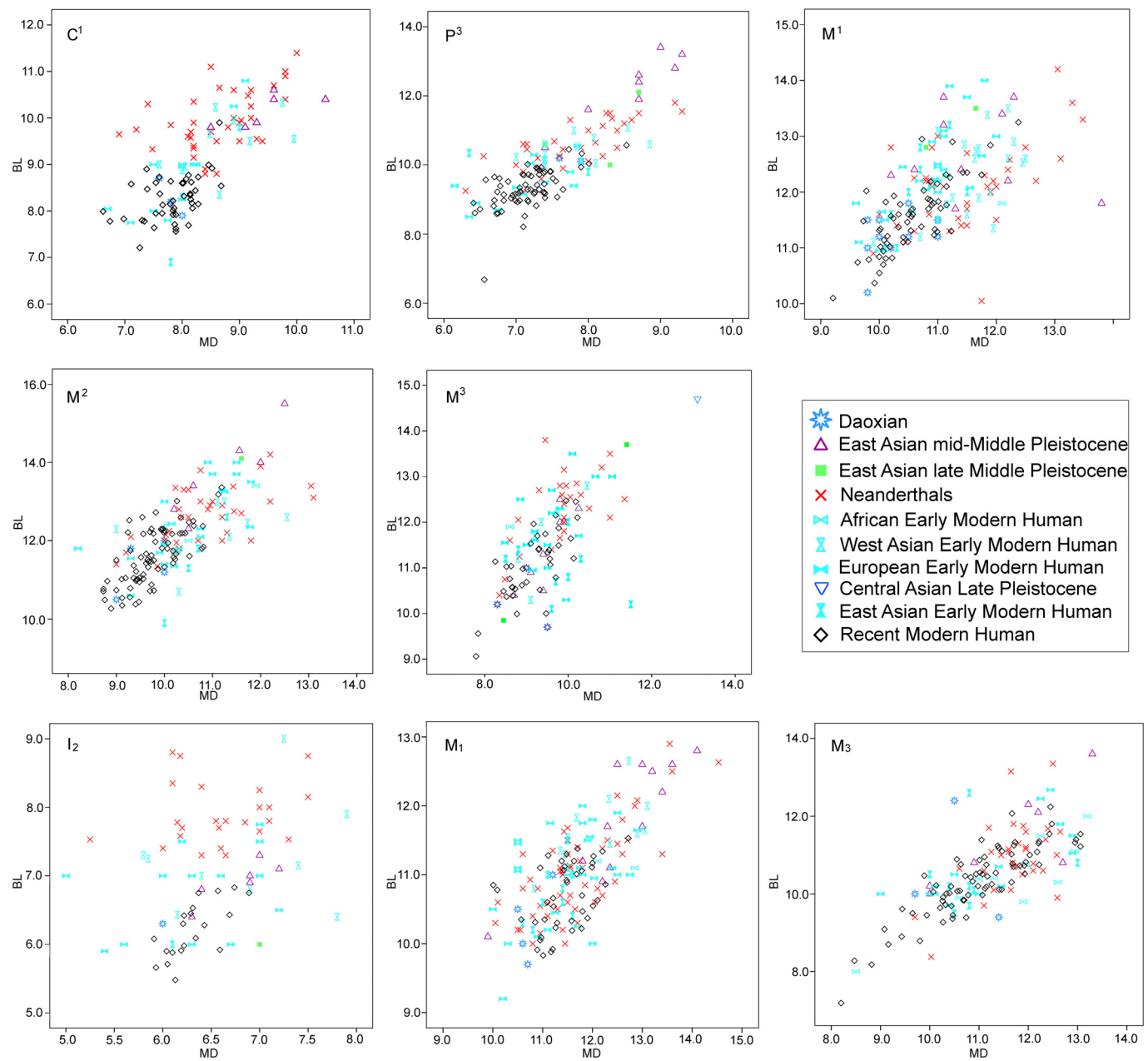


(f)

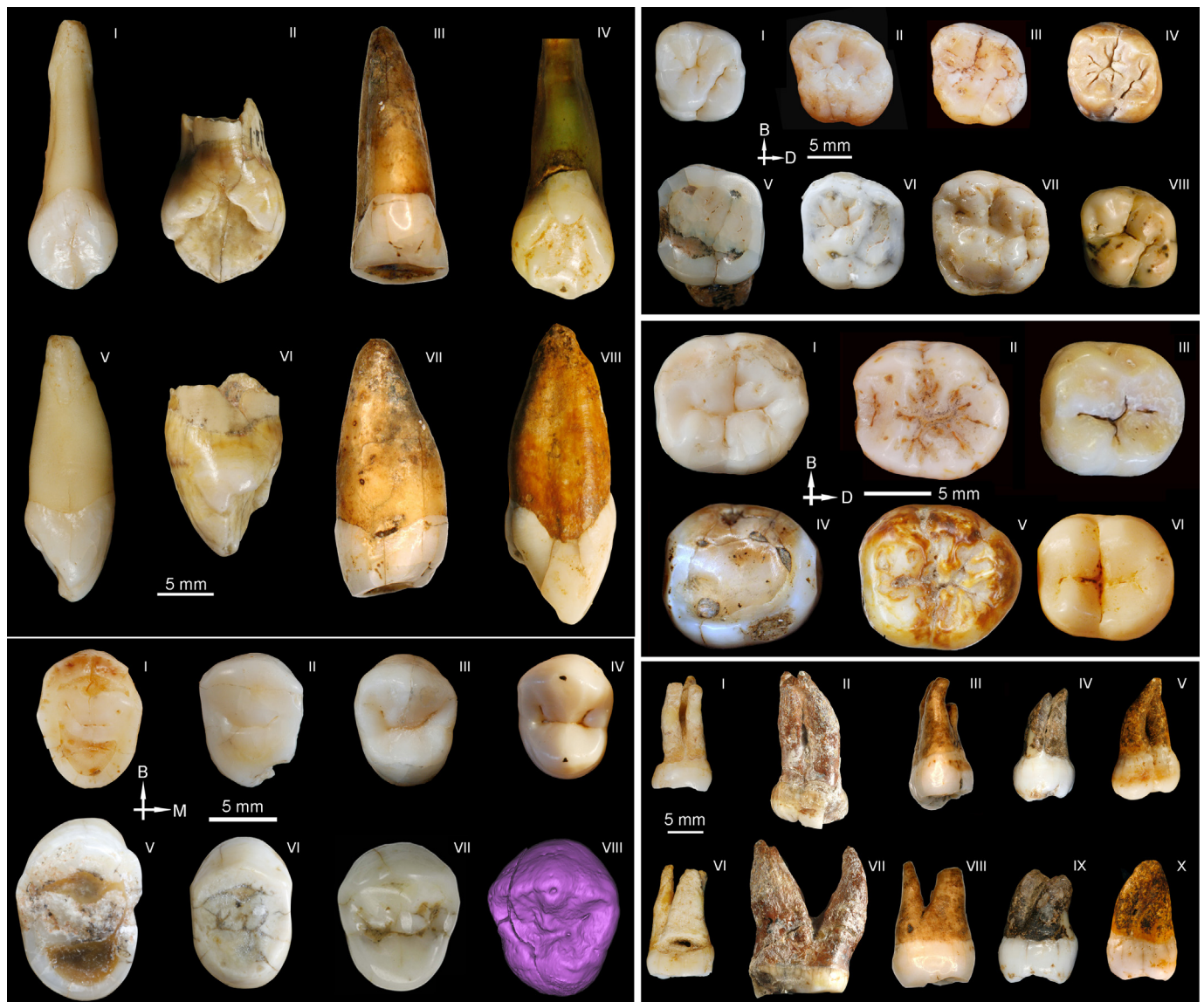
ID#	Demag type	DEC	INC	MAD	Q	AF/Tinf	AF/Tsup	NA/A	VGP λ	VGP Φ
D1A	AF	328	41	9.7	1	15 mT	100 mT	NA	57.6	356.2 E
D1D	Hybrid	311	37	10.7	2	15 mT	100 mT	A	43.8	9.1 E
D1E	TH	335	57	8.1	2	400 °C	590 °C	A	70.4	16 E
D2C	Hybrid	91	33	1.2	2	15 mT	70 mT	A	11.2	186.8 E
D2D	AF	101	31	3.9	1	15 mT	100 mT	NA	3.3	181.3 E
D2E	TH	94	37	4.1	2	250 °C	590 °C	A	10.8	182.9 E

Extended Data Figure 4 | Palaeomagnetic and rock-magnetic analysis of Daoxian flowstone. **a**, Location of the orientated handsamples. White arrow indicates sample D1, black arrow indicates sample D2. **b**, Zijdeveld diagram of alternating field demagnetized specimen D1A. Numbers next to the graph denote alternating field step in mT. **c**, Isothermal remanent magnetization (IRM) acquisition curve up to 1T for specimens D1A and D2D. **d**, Progressive stepwise thermal demagnetization of an IRM up to 1T of specimen D1A. **e**, Projection of virtual geomagnetic pole (VGP) of sample D1 with associated α_{95} . **f**, Summary table of the thermal (TH), alternating field (AF), and hybrid

(both AF and TH) palaeomagnetic results. ID# denotes sample identification. A, anchored; DEC, declination of characteristic remanent magnetization (ChRM) direction; demag, demagnetization; INC, inclination of ChRM direction; MAD, maximum angular deviation; NA, not anchored; NRM, natural remanent magnetization; Q, quality index of ChRM direction, with 1 the highest quality and 2 the lowest; VGP, virtual geomagnetic pole latitude. AF/Tinf, lowest AF level or temperature step of ChRM in mT or °C; AF/Tsup, highest AF level or temperature step of ChRM in mT or °C.



Extended Data Figure 5 | Metric comparison of Daoxian teeth. Bivariate plots of the mesiodistal (MD) and buccolingual (BL) diameters of C¹, P³, M¹, M², M³, I₂, M₁ and M₃ of Daoxian and comparative samples.



Extended Data Figure 6 | Morphological comparison of Daoxian teeth.

Comparative morphology of the Daoxian human teeth with other Pleistocene hominins and modern humans. Top left, upper canines. I, V: Daoxian (DX37); II, VI: Xujiayao (PA1480); III, VII: Huanglong Cave; IV, VIII: modern human. Bottom left: upper third premolars. I, II, III: Daoxian (DX13, DX 29, DX42); IV: modern human; V: Chaoxian; VI: Changyang (PA76); VII: Panxian Dadong (PA1577); VIII: Xujiayao (PA1480). Top right, upper first

molars. I: Daoxian (DX28); II: Neanderthal (Petit-Puymoyen Mx6); III: Qafzeh 5; IV: Tubo (PA1471); V: Hexian (PA836); VI: Chaoxian; VII: Xujiayao (PA1480); VIII: modern human. Middle right, lower second molars. I: Daoxian (DX30); II: Neanderthal (Hortus IV); III: Dolni Vestonice (DV37); IV: Huanglong Cave; V: Xintai; VI: modern human. Bottom right, upper third molars. I, IV: Daoxian (DX17); II, VII: Xujiayao; III, VIII: Huanglong Cave; IV, IX: Tubo (PA1476); V, X: modern humans.

Extended Data Table 1 | List of faunal composition at Daoxian and other Late Pleistocene localities of southern China

Fuyan Cave (Daoxian) ³¹	Huanglong Cave ³²	Zhiren Cave ³³	Liujiang ^{34,35}
<i>Anourosorex squamipes</i>	+		
<i>Soriculus</i> sp.	<i>S. leucops</i>	+	
<i>Rhinolophus ferrumequinum</i>	+	<i>R. pearsoni</i>	
<i>Hipposideros armiger</i>	+	<i>H. pratti</i>	
<i>Eptesicus serotinus</i>			
<i>Murina leucogaster</i>	+		
<i>Tadarida insignis</i>			
<i>Trachypithecus</i> sp.	+	+	
<i>Macaca</i> sp.	<i>M. mulatta</i>	+	
<i>Hylobates</i> sp.	+	+	<i>Pongo</i> sp.
Pteromyidae indet.	<i>Belomys</i>	<i>Petaurista</i>	
<i>Rhizomys</i> sp.	+		
<i>Rattus norvegicus</i>		+	<i>R. rattus</i>
<i>Leopoldamys edwardsi</i>	+	+	
<i>Hystrix subcristata</i>	+	+	sp.
<i>Cuon javanicus</i>	+		sp.
<i>Ursus thibetanus</i>	+	+	sp.
<i>Ailuropoda baconi</i>*	+		<i>A. melanoleucus</i>
<i>Martes flavigula</i>			
<i>Arctonyx collaris</i>	+	+	
<i>Lutra lutra</i>	+		
<i>Viverricula</i> sp.			
<i>Viverra</i> sp.	<i>V. zibetha</i>	+	
<i>Crocota ultima</i>	+		
<i>Panthera pardus</i>		+	+
<i>Panthera tigris</i>	+		sp.
<i>Prionailurus bengalensis</i>	+	<i>Felis</i> sp.	
<i>Stegodon orientalis</i>	+		+
<i>Elephas maximus</i>		+	
<i>Megatapirus augustus</i>	+	+	+
<i>Dicerorhinus sumatrensis</i>	<i>D. kirchbergensis</i>	<i>Rh. sinensis</i>	<i>Rh. sinensis</i>
<i>Sus</i> sp.	<i>S. xiaozhu</i>	<i>S. cf. xiaozhu</i>	
<i>Sus scrofa</i>	+	+	sp.
<i>Moschus</i> sp.	<i>M. moschiferus</i>		
<i>Muntiacus muntjak</i>	+	sp.	sp.
<i>Cervus nippon</i>			sp.
<i>Cervus unicolor</i>	+	+	
<i>Capricornis sumatraensis</i>	+		
<i>Bos (Bibos) gaurus</i>	<i>Bubalus</i>		Bovidae indet.
Number of extinct large mammal species: 5	Number of extinct large mammal species: 9	Number of extinct large mammal species: 5	Number of extinct large mammal species: 3
Percentage of extinct species of large mammals: 19%	Percentage of extinct species of large mammals: 25%	Percentage of extinct species of large mammals: 25%	Percentage of extinct species of large mammals: 23%

Extinct species are marked in bold. References 31–35 are cited in the table.

Extended Data Table 2 | List and measurements of Daoxian teeth

Specimen Number		Tooth class	Side	Occlusal Wear degree	MD	BL	Date of discovery	Location and stratigraphic provenience
Field	Museum							
DX1	PA1543	M ¹	L	3	10.5	11.8	October 8, 2011	IIA Out of context
DX2	PA1544	C ¹	L	5	7.8	8.2	October 8, 2011	IIA Out of context
DX3	PA1545	P ₁	R	4	8.0	8.1	October 10, 2011	IIA Layer 2
DX4	PA1546	M ¹	L	4	(9.8)	10.2	October 10, 2011	IIA Layer 2
DX5	PA1547	M ¹	R	4	11.0	11.5	October 10, 2011	IIA Layer 2
DX6	PA1548	M ¹	R	4	10.2	11.0	October 11, 2011	IIA Out of context
DX7	PA1549	C ¹	R	5	7.6	8.7	October 13, 2011	IIA Layer 2
DX8	PA1550	M ¹	L	5	(10.0)	11.5	September 2, 2012	IIB Layer 2
DX9	PA1551	M ₂	L	6	11.0	10.5	September 3, 2012	IIB Layer 2
DX10	PA1552	M ₃	L	3	10.5	12.4	September 3, 2012	IIB Layer 2
DX11	PA1553	I ₂	L	4	6.0	6.3	September 3, 2012	IIB Layer 2
DX12	PA1554	M ²	L	3	10.0	11.2	September 11, 2012	IIB Layer 2
DX13	PA1555	P ³	L	4	7.6	10.2	September 11, 2012	IIB Layer 2
DX14	PA1556	M ²	R	3	9.0	10.5	September 12, 2012	IIB Layer 2
DX15	PA1557	M ₁	L	4	11.2	11.0	September 19, 2012	IIB Layer 2
DX16	PA1558	M ¹	L	5	9.8	11.0	September 21, 2012	IIB Layer 2
DX17	PA1559	M ³	L	6	8.3	10.2	September 22, 2012	IIB Layer 2
DX18	PA1560	C ₁	R	2	6.9	6.6	September 22, 2012	IIB Layer 2
DX19	PA1561	M ₂	L	6	10.5	10.3	September 23, 2012	IIC Layer 2
DX20	PA1562	M ¹	L	2	11.0	11.5	September 23, 2012	IIC Layer 2
DX21	PA1563	M ³	R	0	9.5	9.7	September 23, 2012	IIC Layer 2
DX22	PA1564	M ₃	L	3	9.7	10.0	September 23, 2012	IIC Layer 2
DX23	PA1565	C ₁	L	2	6.8	7.4	September 23, 2012	IIC Layer 2
DX24	PA1566	M ¹	R	1	10.5	11.2	September 24, 2012	IIC Layer 2
DX25	PA1567	M ₁	R	6	10.7	9.7	September 24, 2012	IIC Layer 2
DX26	PA1568	P ⁴	R	6	6.5	8.5	September 24, 2012	IIC Layer 2
DX27	PA1569	M ₃	R	3	11.4	9.4	September 24, 2012	IIC Layer 2
DX28	PA1570	M ¹	L	1	10.0	11.2	September 24, 2012	IIC Layer 2
DX29	PA1571	P ³	L	3	7.9	10.1	September 24, 2012	IIC Layer 2
DX30	PA1581	M ₃	R	3	11.0	10.5	November 22, 2013	IIC Layer 2
DX31	PA1582	M ¹	L	6	(9.8)	11.5	November 22, 2013	IIC Layer 2
DX32	PA1583	dm ²	R	5	9.8	10.4	November 22, 2013	IIC Layer 2
DX33	PA1584	dm ²	L	4	8.6	10.2	November 22, 2013	IIC Layer 2
DX34	PA1585	P ₁	L	5	7.1	7.5	November 22, 2013	IIC Layer 2
DX35	PA1586	M ¹	R	5	11.0	11.2	November 22, 2013	IIC Layer 2
DX36	PA1587	M ²	L	2	10.0	12.0	November 23, 2013	IIC Layer 2
DX37	PA1588	C ¹	R	2	8.0	7.9	November 23, 2013	IIC Layer 2
DX38	PA1589	P ⁴	R	6	6.2	9.8	November 24, 2013	IIC Layer 2
DX39	PA1590	M ³	R	2	9.0	11.0	November 24, 2013	IIC Layer 2
DX40	PA1591	P ⁴	R	6	6.2	9.1	November 25, 2013	IIC Layer 2
DX41	PA1592	M ²	L	6	9.3	11.8	November 26, 2013	IIC Layer 2
DX42	PA1593	P ³	L	1	7.3	9.9	November 26, 2013	IIC Layer 2
DX43	PA1594	P ⁴	R	6	-	9.0	November 26, 2013	IIC Layer 2
DX44	PA1595	M ₁	L	5	10.5	10.5	November 26, 2013	IIC Layer 2
DX45	PA1596	M ₁	R	5	10.6	10.0	November 26, 2013	IIC Layer 2
DX46	PA1597	M ₂	R	7	-	10.0	November 26, 2013	IIC Layer 2
DX47	PA1598	M ¹	R	5	11.0	11.5	November 27, 2013	IIC Layer 2

List of the Daoxian dental remains by tooth class with the degree of occlusal wear (following ref. 36) crown measurements, and region and stratigraphic position. L, left; R, right. Measurements are given in millimetres.

Extended Data Table 3 | Comparative material

Geography/ Chronology	Specimens	Sources of metrics
Africa		
Late Pleistocene	Herto, Klasies River Mouth*, Mumba	37-39
Holocene	Mesolithic North African sample* (Afalou, Tebessa, Aïn Meterchem, Gambetta, Aïn Dokkara, Taforalt)	
East Asia		
Mid-Middle Pleistocene	Chenjiawo*, Hexian*, Yiyuan*, Zhoukoudian ZKD)	12,30
Late-Middle Pleistocene	Changyang, Chaoxian*, Dingcun*, Jinniushan, Panxian Dadong*, Tongzi*, Xujiayao*, Zhoukoudian Locality 4	40-44
Late Pleistocene	Bailian Cave, Baojiyan, Changwu, Chuandong, Duan, Huanglong Cave*, Huli Cave, Jimuyan, Lipu, Liujiang*, Longlin Longdong, Longtanshan, Luna Cave, Nanshan Cave, Tiandong, Tianyuan Cave*, Tubo*, Xichou, Xintai*, Zhaotong, Zhiren Cave*, Upper Cave	45-60
Holocene and contemporary modern humans	Henan Province, Hubei Province	---
Central Asia		
Late Pleistocene	Denisova	27
West Asia		
Late Pleistocene	Qafzeh*, Skhul	Contributed by Wolpoff
Holocene and contemporary modern humans	Eynan*, Hayonim*, Nahal Oren*, Ohalo*	--
Neanderthals	Amud*, Tabun*, Kebara*, Shanidar	--
Europe		
Neanderthals	Arcy Grotte Renne*, Arcy Hyene*, Arcy Sur Cure (Mousterian), Chateaufort, Ehringsdorf, Genay (Côte d'Or), Gibraltar, Hortus, Krapina, Kulna, La Chaise, La Ferrassie, La Quina, Monsempron*, Le Moustier, Ochoz, Pech de l'aze, Petit Puymoyen, Regourdou, Saccopastore, Sakajia, Spy, St. Césaire, Subalyuk, Vindija	Contributed by Wolpoff
Late Pleistocene	Abri Pataud*, Brno, Combe Capelle, Dolní Věstonice*, Cro-Magnon, Fontchevade, Isturitz*, Le Rois*, Les Vachons, Mladeč, Pavlov, Predmostí, Saint Germain-La Riviere*, Zlatý Kun	Contributed by Wolpoff and ⁶¹
Holocene and contemporary modern humans	Hispanic-muslim medieval collection of San Nicolás (Murcia, Spain)*, Mesolithic French sample* (Tévéc and Hoëdic), Neolithic French sample* (Avize, Dolmens de Bretons, Caverne de L'Homme Mort, Orrouy)	

Detailed list of the samples included in the morphological and metric comparison. Asterisk indicates that we examined the original fossil. For the rest, we used high resolution casts. References 37–61 are cited in the table.

Extended Data Table 4 | Upper first molar relative cusp and occlusal polygon areas

Samples	Protocone		Paracone		Metacone		Hypocone		Polygon	
	N	X±SD	N	X±SD	n	X±SD	n	X±SD	n	X±SD
Daoxian	6	32.5±1.0	6	25.9±1.4	6	22.3±1.3	6	20.3±1.3	4	34.0±1.8
Modern humans	50	30.9±1.1	50	27.0±1.4	50	21.8±1.5	50	20.3±1.6	24	37.5±5.4
Neanderthal	21	29.9±2.4	21	25.8±2.1	21	20.6±1.8	21	23.7±2.1	17	26.7±1.8
Qafzeh	7	31.3±2.3	7	24.8±1.6	7	21.3±2.5	7	22.8±5	4	33.3±2.7
LP HSAP	15	31.8±1.5	15	25.7±2.3	15	22.4±1.7	15	20.1±3	5	32.7±1.9

Data for Qafzeh and Late Pleistocene *H. sapiens* are taken from refs 62 and 63. Late Pleistocene *H. sapiens* (LP HSAP) sample is composed by Dolni Vestonice, Fontchevade, Laugerie Basse, Les Rois, Madeleine, Mladec, Patud, St Germaine-la-Rivière and Vachons.

Telomerase activation by genomic rearrangements in high-risk neuroblastoma

Martin Peifer^{1,2*}, Falk Hertwig^{2,3*}, Frederik Roels^{2,3*}, Daniel Dreidax^{4*}, Moritz Gartlgruber^{4*}, Roopika Menon^{5,6}, Andrea Krämer^{2,3}, Justin L. Roncalioli⁷, Frederik Sand², Johannes M. Heuckmann⁶, Fakhera Ikram^{2,3,8}, Rene Schmidt⁹, Sandra Ackermann^{2,3}, Anne Engesser³, Yvonne Kahlert³, Wenzel Vogel⁵, Janine Altmüller⁸, Peter Nürnberg^{2,8,10}, Jean Thierry-Mieg¹¹, Danielle Thierry-Mieg¹¹, Aruljothi Mariappan², Stefanie Heynck⁶, Erika Mariotti⁶, Kai-Oliver Henrich⁴, Christian Gloeckner⁶, Graziella Bosco¹, Ivo Leuschner¹², Michal R. Schweiger¹³, Larissa Savelyeva⁴, Simon C. Watkins¹⁴, Chunxuan Shao¹⁵, Emma Bell⁴, Thomas Höfer¹⁵, Viktor Achter¹⁶, Ulrich Lang^{16,17}, Jessica Theissen³, Ruth Volland³, Maral Saadati¹⁸, Angelika Eggert¹⁹, Bram de Wilde²⁰, Frank Berthold³, Zhiyu Peng²¹, Chen Zhao²², Leming Shi²², Monika Ortmann²³, Reinhard Büttner²³, Sven Perner⁵, Barbara Hero³, Alexander Schramm²⁴, Johannes H. Schulte^{19,25,26}, Carl Herrmann^{27,28,29}, Roderick J. O'Sullivan⁷, Frank Westermann^{4*}, Roman K. Thomas^{1,23*} & Matthias Fischer^{2,3,30*}

Neuroblastoma is a malignant paediatric tumour of the sympathetic nervous system¹. Roughly half of these tumours regress spontaneously or are cured by limited therapy. By contrast, high-risk neuroblastomas have an unfavourable clinical course despite intensive multimodal treatment, and their molecular basis has remained largely elusive^{2–4}. Here we have performed whole-genome sequencing of 56 neuroblastomas (high-risk, $n = 39$; low-risk, $n = 17$) and discovered recurrent genomic rearrangements affecting a chromosomal region at 5p15.33 proximal of the telomerase reverse transcriptase gene (*TERT*). These rearrangements occurred only in high-risk neuroblastomas (12/39, 31%) in a mutually exclusive fashion with *MYCN* amplifications and *ATRX* mutations, which are known genetic events in this tumour type^{1,2,5}. In an extended case series ($n = 217$), *TERT* rearrangements defined a subgroup of high-risk tumours with particularly poor outcome. Despite a large structural diversity of these rearrangements, they all induced massive transcriptional upregulation of *TERT*. In the remaining high-risk tumours, *TERT* expression was also elevated in *MYCN*-amplified tumours, whereas alternative lengthening of telomeres was present in neuroblastomas without *TERT* or *MYCN* alterations, suggesting that telomere lengthening represents a central mechanism defining this subtype. The 5p15.33 rearrangements juxtapose the *TERT* coding sequence to strong enhancer elements, resulting in massive chromatin remodelling and DNA methylation of the affected region. Supporting a functional role of *TERT*, neuroblastoma cell lines bearing rearrangements or amplified *MYCN* exhibited both upregulated *TERT* expression and enzymatic telomerase activity. In summary, our findings show that remodelling of the genomic context abrogates transcriptional silencing of *TERT* in high-risk

neuroblastoma and places telomerase activation in the centre of transformation in a large fraction of these tumours.

Several sequencing studies have been performed to uncover genomic alterations underlying the diverse clinical phenotypes of neuroblastoma^{2–4}. Neuroblastoma has a low mutation frequency and a heterogeneous mutation spectrum with few recurrently mutated genes. Besides *MYCN*, which is amplified in 20% of neuroblastomas¹, the most frequently mutated gene is the tyrosine kinase gene *ALK*, altered in 8–10% of the cases⁶. Furthermore, inactivating *ATRX* mutations were found in a small subgroup of neuroblastomas with unfavourable outcome^{2,5}. Beyond these alterations, no mechanisms have so far been identified that define high-risk neuroblastoma at the molecular level.

We hypothesized that additional structural alterations might occur in high-risk neuroblastoma and performed whole-genome sequencing of 56 tumours and matched normal controls (Supplementary Table 1). By searching for breakpoint clusters occurring within 100 kilobase-pair (kb) regions⁷ we identified four locations exhibiting clustered breakpoints in more than three samples (Fig. 1a). Three of these regions were related to known genetic alterations in neuroblastoma, namely, *MYCN* amplifications, *ATRX* deletions, and copy number gains of chromosome 17q. The fourth region was located at chromosome 5p15.33 and affected 12 out of 56 tumours (21%, Fig. 1a and Extended Data Fig. 1). Chromosomal regions translocated to the 5p15.33 breakpoints were scattered across chromosome 5 (seven cases) and other chromosomes (five cases). The types of structural alterations were diverse and included balanced rearrangements, translocations associated with single copy number gain, and focal, high-level amplifications (Fig. 1b). Furthermore, the rearrangement at 5p15.33 was caused by chromothripsis affecting chromosome 5 in

¹Department of Translational Genomics, Center of Integrated Oncology Cologne–Bonn, Medical Faculty, University of Cologne, 50931 Cologne, Germany. ²Center for Molecular Medicine Cologne (CMCC), University of Cologne, 50931 Cologne, Germany. ³Department of Pediatric Oncology and Hematology, University Children's Hospital of Cologne, Medical Faculty, University of Cologne, 50937 Cologne, Germany. ⁴Division Neuroblastoma Genomics (B087), German Cancer Research Center, 69120 Heidelberg, Germany. ⁵Department of Prostate Cancer Research, Institute of Pathology, Center for Integrated Oncology Cologne–Bonn, University Hospital of Bonn, 53127 Bonn, Germany. ⁶NEO New Oncology AG, 51105 Cologne, Germany. ⁷Department of Pharmacology and Chemical Biology, University of Pittsburgh Cancer Institute (UPCI), Hillman Cancer Center, Pittsburgh, Pennsylvania 15213, USA. ⁸Cologne Center for Genomics, University of Cologne, 50931 Cologne, Germany. ⁹Institute of Biostatistics and Clinical Research, University of Münster, 48149 Münster, Germany. ¹⁰Cologne Excellence Cluster on Cellular Stress Responses in Aging-Associated Diseases (CECAD), University of Cologne, 50931 Cologne, Germany. ¹¹National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA. ¹²Department of Pathology, University of Kiel, 24118 Kiel, Germany. ¹³Functional Epigenomics, University of Cologne, 50931 Cologne, Germany. ¹⁴Department of Cell Biology, Center for Biologic Imaging, University of Pittsburgh, Pittsburgh, Pennsylvania 15261, USA. ¹⁵Division of Theoretical Systems Biology, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany. ¹⁶Computing Center, University of Cologne, 50931 Cologne, Germany. ¹⁷Department of Informatics, University of Cologne, 50931 Cologne, Germany. ¹⁸Division of Biostatistics, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany. ¹⁹Department of Pediatric Oncology and Hematology, Charité University Medical Center Berlin, 10117 Berlin, Germany. ²⁰Center for Medical Genetics, Ghent University, 9000 Ghent, Belgium. ²¹BGI-Shenzhen, Bei Shan Industrial Zone, Yantian District, Shenzhen, Guangdong, 518083 China. ²²Center for Pharmacogenomics and Fudan-Zhangjiang Center for Clinical Genomics, State Key Laboratory of Genetic Engineering and MOE Key Laboratory of Contemporary Anthropology School of Pharmacy and School of Life Sciences, Fudan University, Shanghai 201203, China. ²³Department of Pathology, University of Cologne, 50937 Cologne, Germany. ²⁴Department of Pediatric Oncology and Hematology, University Children's Hospital, 45147 Essen, Germany. ²⁵German Cancer Consortium (DKTK), 10117 Berlin, Germany. ²⁶German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany. ²⁷Institute of Pharmacy and Molecular Biotechnology, University of Heidelberg, 69120 Heidelberg, Germany. ²⁸Bioquant Center, University of Heidelberg, 69120 Heidelberg, Germany. ²⁹Division of Theoretical Bioinformatics, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany. ³⁰Max Planck Institute for Metabolism Research, 50931 Cologne, Germany.

*These authors contributed equally to this work.

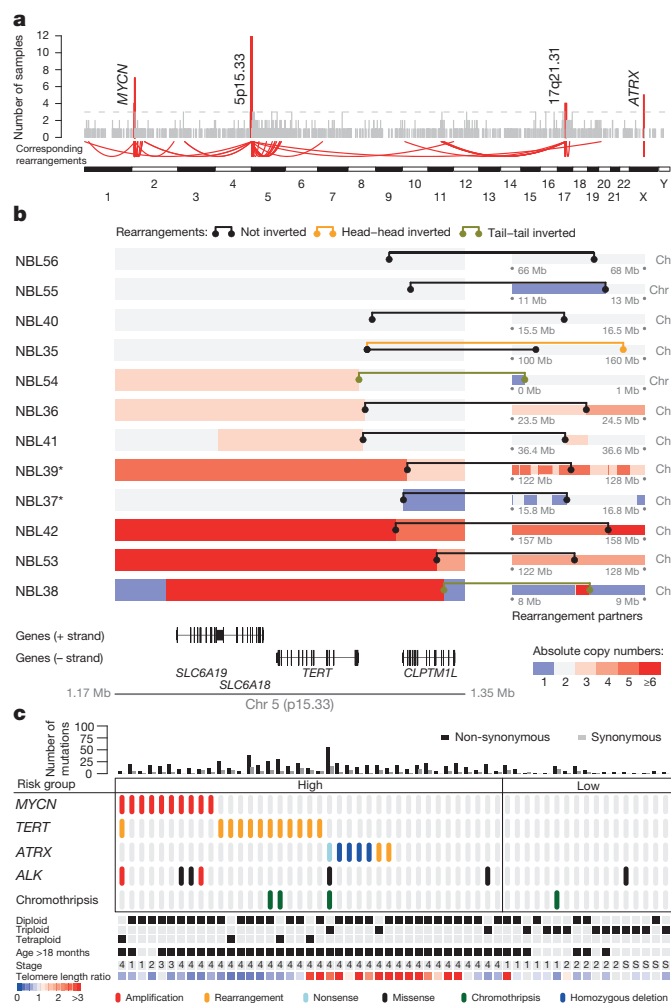


Figure 1 | Genomic rearrangements are clustered at chromosome 5p15.33 in high-risk neuroblastoma. **a**, Distribution of genomic rearrangements occurring within regions of 100 kb in 56 primary neuroblastomas. Rearrangements clustering in more than three tumours are highlighted in red. **b**, Detail of genomic translocations occurring at chromosome 5p15.33 ($n = 12$) and their corresponding rearrangement partner (right). Levels of genomic copy numbers are colour-coded. **c**, Prevalence of *MYCN* amplification, *TERT* rearrangements, genomic alterations of *ATRX* and *ALK*, and chromothripsis in 56 primary neuroblastomas. Samples are ordered from left to right. The number of somatic mutations per tumour and the clinical risk group assessment are given at the top. Tumour ploidy, telomere length ratio (both estimated from sequencing data), age of patient at diagnosis, and tumour stage are displayed at the bottom. S, stage 4S tumours.

two tumours (Fig. 1b and Extended Data Fig. 2a). We noticed that the rearrangements in the 5p15.33 region consistently clustered in a region 50 kb upstream of the *TERT* transcriptional start site without directly affecting the gene or its core promoter region (Fig. 1b)^{8,9}. We did not observe mutations affecting *TERT* itself or its promoter^{10,11}.

In accordance with previous studies^{2,3}, we found an overall low rate of non-synonymous mutations with 13.3 mutations per genome on average (Fig. 1c and Supplementary Table 2), and a significantly lower mutation rate in low-risk tumours than in the high-risk group (low-risk, 5.9 ± 5.5 mutations; high-risk, 16.6 ± 9.9 mutations; $P < 0.001$). Four cases exhibited chromothripsis affecting chromosomes 5, 17, and 20 (Extended Data Fig. 2). We detected only three genes that were expressed and altered in more than two samples (*MYCN*, $n = 10$; *ALK*, $n = 7$; and *ATRX*, $n = 7$). Notably, *TERT* rearrangements occurred exclusively in high-risk tumours in 12 out of 39 of the cases (31%, $P = 0.01$), similar to *ATRX* mutations (Fig. 1c). By contrast, *ALK*

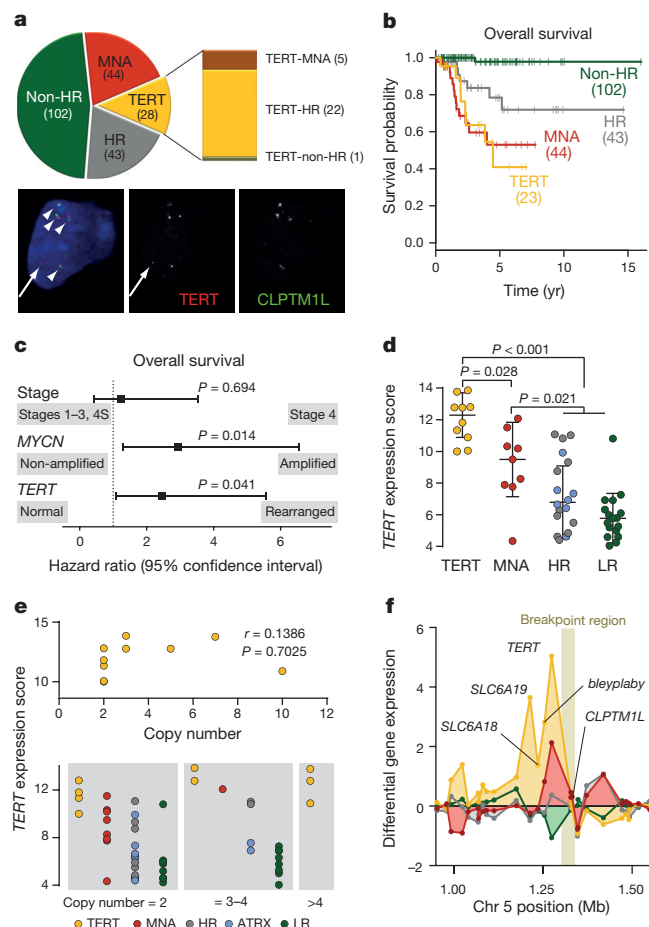


Figure 2 | Genomic *TERT* rearrangements are associated with poor patient outcome and high *TERT* mRNA expression. **a**, Prevalence of *TERT* rearrangements in 217 primary neuroblastomas. *TERT* rearrangements were identified by whole-genome or targeted sequencing and break-apart FISH, exemplarily shown in the lower panel (red, *TERT*; green, *CLPTM1L*). **b**, Overall survival of neuroblastoma patient groups defined by *TERT* rearrangements (*TERT*), *MYCN* amplification (MNA), high-risk disease without these alterations (HR), and low-risk or intermediate-risk disease (non-HR). Patients with tumours bearing both a *TERT* rearrangement and *MYCN* amplification ($n = 5$) were excluded. Overall survival at 5 years: 0.41 ± 0.16 (*TERT*) versus 0.54 ± 0.10 (MNA) versus 0.79 ± 0.08 (HR) versus 0.98 ± 0.02 (LR). **c**, Multivariable Cox regression analysis of the potential prognostic factors stage, *MYCN* status, and *TERT* status for overall survival in patients aged >18 months ($n = 125$). **d**, Distribution of *TERT* mRNA levels derived from transcriptome sequencing of the groups defined above: tumours with *TERT* rearrangements (yellow, $n = 10$), *MYCN* amplifications (red, $n = 9$), high-risk tumours without the aforementioned aberrations (grey, $n = 18$), among which *ATRX*-mutated cases are highlighted by blue circles ($n = 7$), and low-risk tumours (green, $n = 17$). Error bars, median expression and s.d. **e**, Comparison of *TERT* expression in those tumour subgroups defined by *TERT* copy numbers (bottom: left, two copies; middle, three or four copies; right, more than four copies). *TERT* expression levels in relation to *TERT* copy numbers in *TERT* rearranged cases (top). **f**, Relative average gene expression levels at the *TERT* locus. Subgroups of tumours with *TERT* rearrangements (yellow), *MYCN* amplification (red), high-risk tumours without the aforementioned aberrations (grey), and low-risk tumours (green) are compared with average expression in a large cohort of neuroblastoma samples ($n = 498$).

mutations and chromothripsis were found across both risk groups (Fig. 1c). Furthermore, *TERT* rearrangements, *ATRX* alterations, and *MYCN* amplifications occurred in a mutually exclusive fashion within the high-risk group ($P = 0.008$), suggesting that they may converge on similar effector functions.

To further clarify the clinical relevance of *TERT* rearrangements in a large cohort of patients, we examined 161 additional primary

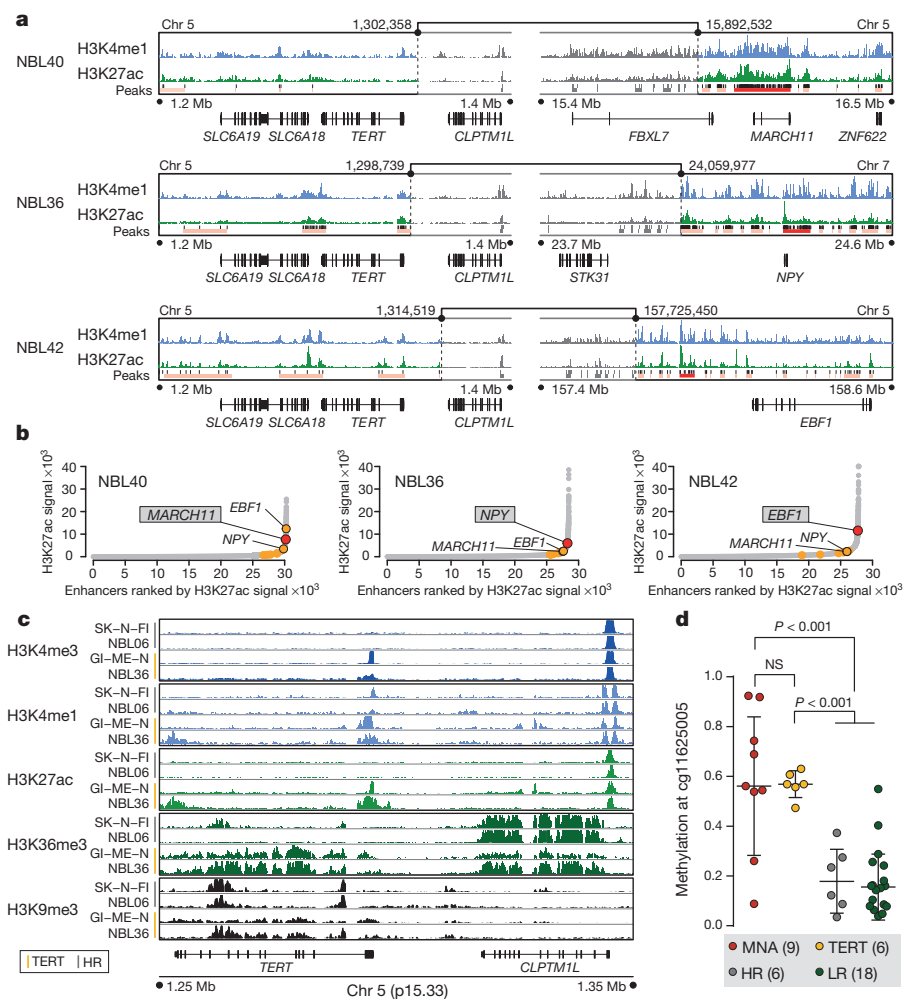


Figure 3 | Translocation of active enhancers drive *TERT* expression in *TERT*-rearranged neuroblastomas. **a**, Normalized read counts of H3K27ac and H3K4me1 histone marks derived from ChIP-seq at the *TERT* rearrangement region in three neuroblastomas. Significant peaks of H3K27ac read counts are displayed by black bars. Enhancer elements identified by stitched peak calls within 12.5 kb regions are shown in pale red. The enhancer element showing highest peak signals within a 0.5 Mb region upstream of the rearrangement breakpoint is highlighted in dark red. **b**, Enhancer elements were ranked according to cumulated read counts over the stitched peak calls. The strongest element identified within the rearranged region of the respective sample is depicted by red circles. Orange circles indicate the strongest enhancer elements in genomic regions affected by rearrangements in other *TERT*-rearranged tumours ($n = 10$; only non-inverted rearrangements are shown). **c**, Survey of the five histone marks H3K4me3, H3K4me1, H3K27ac, H3K36me3, and H3K9me3 in neuroblastomas harbouring *TERT* rearrangements (cell line G1-ME-N and primary tumour NBL36; *TERT*), and neuroblastomas lacking *TERT* and *MYCN* alterations (cell line SK-N-FI and primary tumour NBL06; HR). **d**, DNA methylation of the CpG site cg11625005 proximal to the *TERT* core promoter in *TERT*-rearranged (*TERT*, $n = 6$) and *MYCN*-amplified cases (MNA, $n = 9$) as well as tumours without these alterations (high-risk neuroblastoma, $n = 6$; low-risk neuroblastoma, $n = 18$). Error bars, median methylation and s.d.

neuroblastomas by fluorescence *in situ* hybridization (FISH) and targeted sequencing, and found 16 additional *TERT*-rearranged cases. In the entire set ($n = 217$, Extended Data Table 1a), we detected 28 *TERT* rearrangements (13%), 27 of which occurred in high-risk tumours (Fig. 2a and Extended Data Table 1b). The remaining case was classified as intermediate risk (stage 3, 8 years at diagnosis), and the patient is currently in complete remission after surgery and chemotherapy. *TERT* rearrangements were strongly associated with the unfavourable prognostic markers of stage 4 disease and patient age at diagnosis older than 18 months, but occurred predominantly in tumours without *MYCN* amplification in the high-risk cohort (Extended Data Table 1b). Together, 27 of 114 high-risk neuroblastomas (24%) and 22 of 65 *MYCN*-non-amplified high-risk tumours (34%) harboured *TERT* rearrangements, indicating their frequent occurrence in these groups (Fig. 2a). Patients whose tumours harboured *TERT* rearrangements had a poor clinical outcome, which was similar to that of patients with *MYCN*-amplified tumours and significantly worse than that of high-risk patients without *MYCN* amplification or *TERT* rearrangements (overall survival, $P = 0.056$; Fig. 2b; event-free survival, $P = 0.038$; Extended Data Fig. 3a). The clinical relevance of *TERT* alterations was substantiated by multivariable analyses, in which *TERT* rearrangements predicted unfavourable outcome independently of the established variables stage and *MYCN* (Fig. 2c and Extended Data Fig. 3b).

Since MYCN is a known transcriptional activator of *TERT*¹², we compared *TERT* expression in tumours with 5p15.33 rearrangements with that in *MYCN*-amplified specimens and with the remaining cases. *TERT* expression was significantly higher in neuroblastomas with rearrangements than in *MYCN*-amplified tumours ($P = 0.028$) and

in those without these alterations ($P < 0.001$, Fig. 2d and Extended Data Fig. 3c). The median *TERT* expression was 92-fold higher in *TERT*-rearranged tumours than in low-risk tumours (Extended Data Fig. 3d). In addition, *TERT* expression was higher in *MYCN*-amplified tumours than in the remaining samples ($P = 0.021$, Fig. 2d).

Since the region upstream of *TERT* consists of condensed chromatin in most somatic cells, which keeps *TERT* in a silenced state¹³, we hypothesized that the rearrangements may induce *TERT* transcriptional upregulation by genomic repositioning, rather than by gain of *TERT* copy numbers. In support of this hypothesis, *TERT* expression was massively increased in *TERT*-rearranged tumours compared with low-risk tumours in both subgroups with and without additional *TERT* copies (Fig. 2e). In addition, *TERT* copy numbers and messenger RNA (mRNA) levels did not correlate in *TERT*-rearranged neuroblastomas (Fig. 2e), indicating that additional *TERT* copies have little influence on *TERT* expression in these tumours. In line with that notion, we found evidence for mono-allelic *TERT* expression in five evaluable tumours, suggesting that only the rearranged allele is expressed in these cases (Extended Data Fig. 3e). Additionally, we found no evidence for the existence of tandem duplications affecting *TERT* without the presence of complex rearrangements (Fig. 1b). Together, these findings suggest that remodelling of the genomic context by the rearrangement itself—rather than copy number gain of this region^{14,15}—is the major cause of *TERT* mRNA upregulation in *TERT*-rearranged neuroblastomas.

We also noticed that the distance of the breakpoints to the transcriptional start site of *TERT* did not affect *TERT* expression (Extended Data Fig. 3f). Furthermore, identical rearrangements and elevated *TERT* expression were detected in four tumours biopsied at initial diagnosis

and relapse (Extended Data Fig. 4), supporting the notion that 5p15.33 rearrangements and consecutive upregulation of *TERT* are stable genetic events during neuroblastoma development and progression.

We observed that in *TERT*-rearranged neuroblastomas three genes (*bleyplaby* (AceView-annotated human gene; <http://www.aceview.org>), *SLC6A18*, and *SLC6A19*) neighbouring *TERT* on the distal side of the breakpoint showed markedly increased expression, while expression of *suweeby*, *blorplaby*, and *CLPTMIL* (cleft lip and palate transmembrane protein 1-like protein), located on the proximal side of the breakpoint, did not follow this pattern (Fig. 2f and Extended Data Fig. 5a). By contrast, *MYCN*-amplified tumours exhibited upregulation of only *TERT* and *bleyplaby*, but lacked upregulation of *SLC6A18* and *SLC6A19*, whereas all four genes were silenced in low-risk neuroblastomas and in high-risk tumours without *TERT* rearrangements or *MYCN* amplification (Fig. 2f and Extended Data Fig. 5a, b). Our findings were substantiated by analyses of differential gene expression of tumours bearing *TERT* rearrangements or *MYCN* amplification versus low-risk tumours: in both comparisons, *TERT* was among the most highly upregulated genes, while *SLC6A18* and *SLC6A19* were upregulated in *TERT*-rearranged tumours only (Extended Data Fig. 6a, b). Thus, genomic rearrangements at 5p15.33 abrogate silencing of gene expression in this chromosomal region, whereas amplified *MYCN* selectively upregulates *TERT* by transcriptional activation. In line with this notion, we found that *TERT* was the most strongly downregulated gene upon short-interfering-RNA-mediated *MYCN* knockdown in a *MYCN*-amplified neuroblastoma cell line (Extended Data Fig. 6c). Together, these observations suggest that both 5p15.33 rearrangements and *MYCN* amplification converge on *TERT* activation in high-risk neuroblastoma.

The massive upregulation of *TERT* expression observed in tumours bearing structural rearrangements of the *TERT* locus suggested that translocation of regulatory elements (for example, enhancers) might be responsible for *TERT* activation. We therefore examined patterns of H3K27ac and H3K4me1 histone modifications, which are known to mark active enhancers^{16,17}, in three *TERT*-rearranged tumours using chromatin immunoprecipitation coupled to sequencing (ChIP-seq). In all three tumours, we detected peak signals in the translocated regions that predict the presence of multiple enhancer clusters immediately adjacent to the breakpoints (Fig. 3a). Ranking of H3K27ac peak signals revealed that rearrangements consistently juxtapose *TERT* to strong enhancer elements, several of which are compatible with the recent definition of super-enhancers (Fig. 3b)¹⁸.

To examine the chromatin context of the *TERT* locus further, we analysed additional histone modifications, known to mark active promoters (H3K4me3), transcription elongation (H3K36me3), DNA methylation (H3K9me3), and Polycomb repressive complex 2 (PRC2)-related gene silencing (H3K27me3), in neuroblastoma cells with and without *TERT* rearrangements (Fig. 3c and Extended Data Figs 7 and 8a). While histone modifications indicative of an active promoter (H3K4me3 and H3K27ac) and transcription elongation (H3K36me3) were spread to the *TERT* locus in *TERT*-rearranged cases, these marks were completely lacking in tumours without *TERT* rearrangements (Fig. 3c). By contrast, the promoter and gene body of *TERT* were broadly marked by the repressive mark H3K27me3 in cells lacking *TERT* alterations (Extended Data Fig. 8a).

DNA methylation analyses of 39 primary neuroblastomas revealed a consistent increase in CpG methylation across the *TERT* locus in rearranged and *MYCN*-amplified tumours (Extended Data Fig. 8b). The strongest differential methylation was detected in a CpG site located in close proximity to the core promoter of *TERT* (Fig. 3d and Extended Data Fig. 8b), which has been previously associated with a disabled repressive element and elevated *TERT* expression¹⁹. Thus, structural rearrangements occurring at 5p15.33 lead to juxtaposition of the *TERT* locus to strong enhancer elements, resulting in a massive epigenetic remodelling of the affected region.

Our findings were further supported by analyses of cell lines derived from high-risk neuroblastoma bearing *TERT* rearrangements, *MYCN*

amplification, or none of these alterations (Extended Data Fig. 7). Cell lines with *TERT* translocations or *MYCN* amplification showed elevated *TERT* mRNA expression compared with cell lines lacking such alterations (Fig. 4a). High levels of *TERT* expression were paralleled by a significant increase in enzymatic telomerase activity as assessed by a telomeric repeat amplification assay (Fig. 4b).

ATRX mutations occurred exclusively in tumours lacking *MYCN* amplification and *TERT* rearrangements (Fig. 1c) and have previously been associated with alternative lengthening of telomeres (ALT)⁵. Consistent with the hypothesis that high-risk neuroblastomas may be associated with ALT in the absence of *TERT* activation through *MYCN* amplification or *TERT* rearrangement, the *ATRX*-mutant high-risk neuroblastoma cell line (CHLA-90) has evidence of ALT pathway activation²⁰. Furthermore, we observed activation of the ALT pathway in two high-risk neuroblastoma cell lines without *TERT* or *MYCN* alterations (Extended Data Fig. 9)²¹. In line with these and published observations⁵, we found that *ATRX*-mutated primary neuroblastomas and other high-risk tumours lacking *TERT* or *MYCN* alterations had abundant telomere repeat sequences, pointing towards activation of ALT²² (Figs 1c and 4c). By contrast, telomeres were short in primary

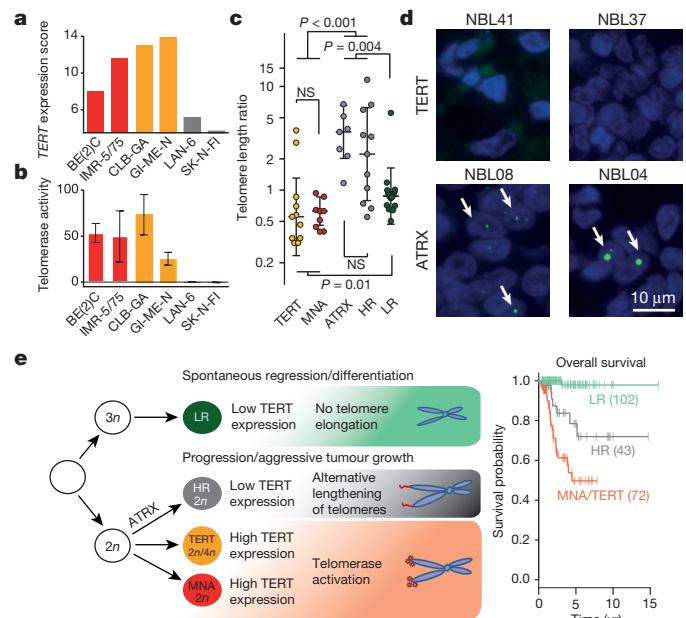


Figure 4 | Telomerase activity is associated with *TERT* rearrangements and *MYCN* amplification, while ALT occurs in high-risk neuroblastoma lacking these alterations. **a**, **b**, *TERT* expression levels as determined by transcriptome sequencing (**a**) and telomerase activity (**b**) as determined by telomeric repeat amplification assay in neuroblastoma cell lines bearing *TERT* rearrangements (GI-ME-N and CLB-GA), *MYCN* amplification (SK-N-BE(2)C and IMR-5/75), and cell lines without these alterations (LAN-6 and SK-N-FI). **c**, Distribution of telomere length ratios between the tumours and matched normals (computed from whole-genome sequencing) in primary neuroblastoma subgroups defined by *TERT*, *MYCN*, and *ATRX* alterations and risk group (HR, high-risk without the aforementioned alterations; LR, low-risk). **d**, Telomere FISH analyses of two *TERT*-rearranged (NBL41, NBL37) and two *ATRX*-mutated (NBL08, NBL04) primary tumours. **e**, A revised model for neuroblastoma pathogenesis based on recurrent genomic alterations, the presence or absence of telomere maintenance pathways, and clinical courses of the disease (modified after ref. 24). In this model, high-risk tumours are distinguished from low-risk tumours by active mechanisms of telomere lengthening. The most aggressive neuroblastomas are defined by telomerase activation as a result of either *TERT* rearrangement (*TERT*) or *MYCN* amplification (MNA). In addition, near-diploid (2n) or near-tetraploid (4n) karyotypes are preferentially observed in high-risk tumours, while near-triploid karyotypes are mostly found in low-risk tumours²⁴. Overall survival of patient subgroups at 5 years: 0.51 ± 0.08 (MNA/*TERT*) versus 0.79 ± 0.08 (high-risk tumours) versus 0.98 ± 0.02 (low-risk tumours).

tumours in which telomerase was activated²³. We confirmed this result by telomere FISH analysis in primary tumours bearing either *ATRX* mutations or *TERT* rearrangements (Fig. 4d). In low-risk tumours, the telomere lengths were similar to matched normal samples, indicating that telomere maintenance mechanisms may be inactive in these tumours (Fig. 4c). Together, these findings support the notion that high *TERT* expression mediated by 5p15.33 rearrangements or *MYCN* amplification induces constitutive catalytic telomerase activity, whereas ALT is activated in high-risk neuroblastoma lacking these alterations. Thus, our results suggest that high-risk neuroblastoma is defined molecularly by mechanisms leading to telomere lengthening.

In summary, we discovered recurrent *TERT* rearrangements in approximately one-quarter of high-risk neuroblastomas. Our results indicate that most high-risk tumours are affected by *TERT* rearrangements, *MYCN* amplification, or *ATRX* mutations, all of which funnel into telomere lengthening, thus providing a molecular, mechanistic definition of this neuroblastoma subtype (Fig. 4e). By contrast, low-risk tumours are characterized by the absence of such alterations and low *TERT* expression levels, presumably leading to the inability to gain immortal proliferation capacity. In our data set, the most aggressive subtype of neuroblastoma was defined by telomerase activation as a result of either *TERT* rearrangement or *MYCN* amplification (Fig. 4e). With the further development of telomerase inhibitors, our finding might point to a novel therapeutic option for the most aggressive subgroup of this deadly paediatric disease.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 17 April; accepted 22 July 2015.

Published online 14 October; corrected online 28 October 2015 (see full-text HTML version for details).

- Maris, J. M., Hogarty, M. D., Bagatell, R. & Cohn, S. L. Neuroblastoma. *Lancet* **369**, 2106–2120 (2007).
- Molenaar, J. J. et al. Sequencing of neuroblastoma identifies chromothripsis and defects in neurogenesis genes. *Nature* **483**, 589–593 (2012).
- Pugh, T. J. et al. The genetic landscape of high-risk neuroblastoma. *Nature Genet.* **45**, 279–284 (2013).
- Sausen, M. et al. Integrated genomic analyses identify *ARID1A* and *ARID1B* alterations in the childhood cancer neuroblastoma. *Nature Genet.* **45**, 12–17 (2013).
- Cheung, N. K. et al. Association of age at diagnosis and genetic mutations in patients with neuroblastoma. *J. Am. Med. Assoc.* **307**, 1062–1071 (2012).
- Mossé, Y. P. et al. Identification of *ALK* as a major familial neuroblastoma predisposition gene. *Nature* **455**, 930–935 (2008).
- Northcott, P. A. et al. Enhancer hijacking activates *GFI1* family oncogenes in medulloblastoma. *Nature* **511**, 428–434 (2014).
- Nagel, I. et al. Deregulation of the telomerase reverse transcriptase (*TERT*) gene by chromosomal translocations in B-cell malignancies. *Blood* **116**, 1317–1320 (2010).
- Davis, C. F. et al. The somatic genomic landscape of chromophobe renal cell carcinoma. *Cancer Cell* **26**, 319–330 (2014).
- Horn, S. et al. *TERT* promoter mutations in familial and sporadic melanoma. *Science* **339**, 959–961 (2013).
- Huang, F. W. et al. Highly recurrent *TERT* promoter mutations in human melanoma. *Science* **339**, 957–959 (2013).
- Mac, S. M., D'Cunha, C. A. & Farnham, P. J. Direct recruitment of N-myc to target gene promoters. *Mol. Cell* **29**, 76–86 (2000).
- Zhao, Y., Wang, S., Popova, E. Y., Grigoryev, S. A. & Zhu, J. Rearrangement of upstream sequences of the *hTERT* gene during cellular immortalization. *Genes Chromosom. Cancer* **48**, 963–974 (2009).
- Kumps, C. et al. Focal DNA copy number changes in neuroblastoma target *MYCN* regulated genes. *PLoS One* **8**, e52321 (2013).
- Cobrinik, D. et al. Recurrent pre-existing and acquired DNA copy number alterations, including focal *TERT* gains, in neuroblastoma central nervous system metastases. *Genes Chromosom. Cancer* **52**, 1150–1166 (2013).
- Heintzman, N. D. et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genet.* **39**, 311–318 (2007).
- Ernst, J. et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
- Whyte, W. A. et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307–319 (2013).
- Castelo-Branco, P. et al. Methylation of the *TERT* promoter and risk stratification of childhood brain tumours: an integrative genomic and molecular study. *Lancet Oncol.* **14**, 534–542 (2013).
- Farooqi, A. S. et al. Alternative lengthening of telomeres in neuroblastoma cell lines is associated with a lack of *MYCN* genomic amplification and with p53 pathway aberrations. *J. Neurooncol.* **119**, 17–26 (2014).
- O'Sullivan, R. J. et al. Rapid induction of alternative lengthening of telomeres by depletion of the histone chaperone ASF1. *Nature Struct. Mol. Biol.* **21**, 167–174 (2014).
- Cesare, A. J. & Reddel, R. R. Alternative lengthening of telomeres: models, mechanisms and implications. *Nature Rev. Genet.* **11**, 319–330 (2010).
- Martinez, P. & Blasco, M. A. Telomeric and extra-telomeric roles for telomerase and the telomere-binding proteins. *Nature Rev. Cancer* **11**, 161–176 (2011).
- Brodeur, G. M. Neuroblastoma: biological insights into a clinical enigma. *Nature Rev. Cancer* **3**, 203–216 (2003).

Supplementary Information is available in the online version of the paper.

Acknowledgements We are indebted to the patients and their parents for making available the tumour specimens analysed in this study. We thank the German Neuroblastoma Biobank for providing samples from patients. The Institutional Review Board approved collection and use of all specimens in this study. We also thank our colleagues N. Hemstedt, H. Düren, E. Hess, J. Kreth, and J. Gopalakrishnan; and our collaboration partners I. Amit and F. Paul at the Weizmann Institute of Science for technical assistance. We further acknowledge the Center for Molecular Medicine Cologne light microscope facility for helping us to obtain high-quality data of FISH analyses, and S. Wolf and the Next Generation Sequencing Unit at the German Cancer Research Center (DKFZ) for sequencing. This work was supported by the German Cancer Aid (grant 110122) to M.F., F.W., J.H.S., A.S., and S.A.; the German Ministry of Science and Education (BMBF) as part of the e:Med initiative (grant 01ZX1303A to M.P., M.F., J.H.S., R.B., U.L., and R.K.T., grant 01ZX1406 to M.P., and grant 01ZX1307D to M.F. and F.W.); the BMBF (grant 0316076A to F.W.); the European Union (grant 259348 to F.W.); the Fördergesellschaft Kinderkrebs-Neuroblastom-Forschung e.V. (to M.F.); the German-Israeli Helmholtz Research School in Cancer Biology (to M.G. and F.W.); the Volkswagenstiftung (Lichtenberg Program) (to M.R.S.); and the Center for Molecular Medicine Cologne.

Author Contributions Conception and design: M.P., F.H., F.R., D.D., M.G., F.W., R.K.T., and M.F. Administrative support, provision of study materials and patients: M.P., J.H., I.L., T.H., P.N., V.A., U.L., A.Eg., F.B., Z.P., C.Z., L.Sh., R.B., S.P., B.H., A.S., J.H.S., F.W., R.J.O., R.K.T., and M.F. Conduct of the experiments, data analysis, and interpretation: M.P., F.H., F.R., D.D., M.G., R.M., A.K., J.L.R., F.S., J.H., F.I., R.S., S.A., A.En., Y.K., W.V., J.A., J.T.-M., D.T.-M., A.M., S.H., E.M., K.-O.H., C.G., G.B., M.-R.S., L.Sa., S.C.W., C.S., E.B., J.T., R.V., M.S., B.D., M.O., B.H., C.H., R.J.O., F.W., R.K.T., and M.F. Manuscript writing: M.P., F.H., F.R., F.W., R.K.T., and M.F. All authors read and approved the final manuscript.

Author Information All high-throughput data have been deposited at the European Genome-phenome Archive (<https://www.ebi.ac.uk/ega/>) under accession number EGAS00001001308. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.F. (matthias.fischer@uk-koeln.de), R.K.T. (roman.thomas@uni-koeln.de), F.W. (frank.westermann@dkfz-heidelberg.de) or M.P. (mpeifer@uni-koeln.de).

METHODS

Patients and tumour samples. Overall, this project comprised tumour samples from 217 German patients with neuroblastoma (Extended Data Table 1a). Patients were diagnosed between 1991 and 2014 and were registered and treated according to several clinical trials of the Gesellschaft für Pädiatrische Onkologie und Hämatologie. The trials were approved by the Ethics Committee of the Medical Faculty, University of Cologne. The Institutional Review Board approved collection and use of all specimens in this study. Informed consent was obtained from all patients. The MYCN gene copy number was determined as a routine diagnostic method using FISH analysis. DNA and total RNA was isolated from tumour samples with at least 60% tumour cell content as evaluated by a pathologist.

TERT rearrangements were established as a novel molecular marker in a discovery cohort of 56 patients. In this set, *TERT* rearrangements ($n = 12$) occurred exclusively in high-risk patients ($n = 39$). We sought to validate this finding in a larger, representative neuroblastoma cohort, comprising approximately 40% high-risk patients. Allowing for a potential occurrence of *TERT* rearrangements in up to 10% of non-high-risk patients and ensuring a statistical power of 80%, we estimated that at least 75 non-high-risk patients were required for validation. We therefore investigated 161 additional tumours derived from 86 non-high-risk and 75 high-risk patients (including 39 and 36 tumours with and without MYCN amplification, respectively). The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Cell lines. Neuroblastoma cell lines LAN-6, GI-ME-N, as well as SK-N-FI were directly purchased from the Deutsche Sammlung von Mikroorganismen und Zellkulturen (DSMZ, Braunschweig, Germany) or American Type Culture Collection (ATCC/LGC Standards, Molsheim Cedex, France), respectively. Furthermore, SK-N-BE(2)C, IMR-5/75, and CLB-GA were provided by the laboratories of H. Deubzer, F. Westermann and J. H. Schulte, respectively. All cell lines not directly purchased from ATCC or DSMZ were authenticated by STR profiling at the DSMZ. IMR-5/75, SK-N-BE(2)C, GI-ME-N, and CLB-GA were grown in RPMI1640 with 10% FBS; LAN-6 and SK-N-FI were cultured in DMEM with 20% FBS. All cell lines were cultured without antibiotics and routinely tested negative for mycoplasma.

Whole-genome sequencing. DNA was extracted from fresh-frozen tumour tissue and the corresponding matched normal of 56 patients using the Puregene Core Kit A (Qiagen) and NucleoSpin Blood DNA extraction kit (Macherey-Nagel), respectively, according to the manufacturers' instructions. DNA was eluted in $1 \times$ TE buffer, diluted to a working concentration of $100 \text{ ng } \mu\text{l}^{-1}$ and stored at -80°C . Libraries were prepared with the TruSeq DNA PCR-free sample preparation kit (Illumina) followed by size selection using SPRI beads (Beckman Coulter Genomics). The final libraries were then sequenced on an Illumina HiSeq 2000 instrument with a paired-end read length of 2×100 nucleotides.

Whole-genome sequencing data analysis. Raw sequencing reads were aligned to the human genome (NCBI build 37/hg19) by using BWA (version 0.6.1-r104; <https://github.com/lh3/bwa>). Possible PCR duplicates were then masked in the resulting alignment files by searching for concordant read pairs. Next, somatic mutations were called using a further development of our in-house mutation caller. The major improvements to previous versions²⁵ of the caller were that identified variants were filtered against a library of more than 500 normals (mixed whole-exome and whole-genome) and that the error model contained contributions of human library contamination which were estimated from the sequencing data. With these modifications we were able to gain a larger sensitivity and specificity (data not shown). Rearrangements as well as copy number changes were analysed as described previously²⁵. Patterns of recurrent genomic rearrangements were identified by scanning the genomes for breakpoint clusters that occur within 100 kb regions in a similar approach to that described in ref. 7. To compute the telomere ratio from whole-genome sequencing data, raw sequencing reads containing the telomere repeat sequence (TTAGGG)_{*n*} or its reverse complement were counted², and the ratio between the tumour and matched normal was determined. This ratio was then normalized to the absolute amount of sequenced DNA using the total amount of reads from the tumour and the normal.

Targeted sequencing. In the validation cohort, hybrid-capture-based target enrichment followed by massively parallel sequencing of the genomic region encompassing *TERT* and *CLPTMIL* was used to detect *TERT* rearrangements. Alignment and the detection of genomic rearrangements were performed analogous to whole-genome sequencing. We reported only those *TERT* rearrangements that had been detected both by FISH analysis and by targeted sequencing.

RNA sequencing. RNA sequencing and gene expression analysis was performed as described previously²⁶. Briefly, a Dynabeads mRNA Purification Kit (Invitrogen) was used to purify mRNA from total RNA. Library construction was performed according to the standard TruSeq protocol. Clusters were generated according to the TruSeq PE cluster Kit version 3 reagent preparation guide (for cBot-HiSeq/HiScanSQ). High-throughput shotgun sequencing was

performed on the IlluminaHiSeq 2000 platform. Paired-end reads with lengths of 100 nucleotides were generated. Raw data processing, read mapping, and gene expression quantification were done using the Magic-AceView analysis pipeline as described²⁷. The Magic analysis tool is accessible at <ftp://ftp.ncbi.nlm.nih.gov/repository/acedb/Software/Magic>; AceView served as primary transcriptome reference (<http://www.aceview.org>). RNA sequencing was also used to identify MYCN-regulated genes in IMR5/75 cells expressing MYCN shRNA under the control of the tet repressor. Briefly, for IMR5/75 shRNA inducible cells, $1 \mu\text{g ml}^{-1}$ tetracycline or the equivalent of volume of 70% ethanol was added to the cells, and cells were incubated for 24 h and then harvested for RNA extraction using RNeasy mini kit (Qiagen). Five micrograms of RNA from each sample was processed using the RiboGold kit (Epicentre) to remove rRNA from samples to increase reads from mRNA. The concentration of the resulting RNA was measured using the Qubit RNA assay (Life Technologies). One microgram of RNA was then used to prepare libraries for sequencing using the NEB Ultra directional RNA library prep lit for Illumina (New England Bioscience) according to the manufacturer's instructions.

Microarray. Single-colour gene expression profiles were generated using customized $4 \times 44 \text{ K}$ oligonucleotide microarrays produced by Agilent Technologies. Labelling and hybridization was performed following the manufacturer's protocol. Microarray expression profiles were generated using Agilent's Feature Extraction software (version 9.5.1). Data were normalized using quantile normalization.

Validation of TERT rearrangements by dideoxy sequencing. Rearrangements of the *TERT* locus were validated by dideoxy sequencing in both diagnostic and relapsed tumour samples. Dideoxy sequencing was performed by SeqLab.

FISH. BAC clones CTD-2191M2 (to detect the region proximal of *CLPTMIL*) and CTD-2511M20 (to detect the *TERT/SLC6A18/SLC6A19* loci) were labelled with digoxigenin and biotin, respectively (see also Fig. 3a). Cell line cytospin preparations were pre-treated with $2 \times$ SSC solution at 37°C for 30 min, digested with Digest-All III (dilution 1:2, Invitrogen) at 37°C for 6 min, fixed in 4% formaldehyde, and subsequently dehydrated in a graded ethanol series. FISH probes and human Cot-1 DNA (Life Technologies) in hybridization buffer (50% formamide, 10% dextran sulfate sodium, in $2 \times$ SSC) were co-denatured at 85°C for 4 min and hybridized overnight at 37°C . Post-hybridization washing was done with $0.5 \times$ SSC at 75°C for 5 min, followed by washes in PBS, a blocking step with CAS-block (Life Technologies, with 10% normal goat serum in PBS) and a 1 h post-incubation with streptavidin-Alexa-555 conjugates (1:500, Life Technologies) and anti-digoxigenin-FITC (1:500, Roche), to enable fluorescence detection. After three subsequent washes in PBS, samples were mounted with VectaShield mounting media containing 4',6-diamidino-2-phenylindole dihydrochloride (DAPI; Vectorlabs). Images were acquired using an Olympus Fluoview FV10 scanning confocal microscope system.

Telomeric repeat amplification protocol assay. Telomerase activity in cell lines was determined with a PCR-based telomeric repeat amplification protocol (TRAP) enzyme-linked immunosorbent assay (ELISA) kit (*TeloTAGGG* Telomerase PCR ELISA^{PLUS}, Roche) according to the manufacturer's protocol.

ChIP-seq analysis of histone modifications. Formaldehyde cross-linking of cells, cell lysis, sonication, ChIP procedure, and library preparation were performed as described previously²⁸, starting with approximately 4×10^6 cells (1×10^6 cells per individual immunoprecipitation). Direct cell lysis for each sample was achieved by incubation for 30 min in 950 μl RIPA I on ice (10 mM Tris-HCl pH 8.0, 1 mM EDTA pH 8.0, 140 mM NaCl, 0.2% SDS, 0.1% DOC). Tissue disruption, formaldehyde fixation, and sonication of tumour material were done according to a previously published protocol²⁹. Approximately 30 mg of fresh-frozen tumour tissue was used per individual ChIP-seq experiment. All subsequent steps were performed analogous to cell line experiments. The Bioruptor Plus sonication device (Diagenode) was used for high intensity sonication for 30–60 min each with intervals of 30 s on and 30 s off. Library preparation was performed using the NEBNext Ultra DNA Library Prep Kit (New England Biolabs) according to the manufacturer's protocol. Samples were mixed in equal molar ratios and sequenced on an Illumina sequencing platform.

ChIP-seq data analysis. Single-end reads were aligned to the hg19 genome using Bowtie2 (version 2.1.0). Only uniquely aligned reads were kept. BAM files of aligned reads were further processed using the deepTools suite (<https://github.com/fidellram/deepTools>). Input files were subtracted from the treatment files using the bamCompare tool, applying the SES method for normalization of signal to noise. Resulting signals were normalized to an average $1 \times$ coverage to produce signal (bigWig) files. Peaks were called using the MACS 1.4 tool using default parameters.

DNA methylation profiling. DNA was isolated from snap-frozen neuroblastoma tissue. Genome-wide DNA methylation was assessed using an Infinium HumanMethylation450 BeadChip (Illumina) according to the manufacturer's instructions. Probes were removed on the basis of the following criteria: (1) proportion of non-detectable β values >0.3 ($n = 379$), (2) single nucleotide

polymorphism at or near the targeted CpG site according to R-Forge package IMA (<https://rforge.net/IMA/>, $n = 92,600$), (3) control probes ($n = 65$), and (4) mapping to the X or Y chromosome ($n = 10,351$). Together, 382,182 probes were kept for further analysis. The k -nearest neighbours method was used to impute missing values, and a subset quantile normalization was applied. *TERT*-related CpGs were annotated using the assignGenomeAnnotation program of the HOMER tool suite (<http://homer.salk.edu/homer>).

Telomere restriction-fragment analysis and C-circle assay. Both assays have been performed as described previously²¹. Briefly, genomic DNA was purified and digested with AluI and MboI. For restriction-fragment analysis, 10 µg of digested DNA was electrophoresed on a 0.8% TBE-agarose gel. Subsequently, telomeric DNA was detected by Southern blotting using a [³²P]dATP end-labelled (CCCTAA)₄ oligonucleotide probe. For the C-circle assay, DNA samples (7.5 ng, 10 µl) diluted in ultraclean water were combined with 10 µl BSA (NEB; 0.2 mg ml⁻¹), 0.1% Tween, 0.2 mM each dATP, dGTP, dTTP, and 1 × Φ29 Buffer (NEB) in the presence or absence of 7.5 U ΦDNA polymerase (NEB), incubated at 30 °C for 8 h and then at 65 °C for 20 min. Reaction products were diluted to 100 µl with 2 × SSC and dot-blotted onto a 2 × SSC-soaked nylon membrane. DNA was ultraviolet cross-linked onto the membrane and hybridized with a ³²P-end-labelled (CCCTAA)₄ oligonucleotide probe to detect C-circle amplification products. All blots were washed, exposed to PhosphorImager screens, scanned, and quantified using a Typhoon 9400 PhosphorImager (Amersham, GE Healthcare). Genomic DNA from ALT-positive (U2OS) cells served as positive control and reference for the quantification of C-circles detected in other cell lines.

Immunofluorescence-FISH for APBs. ALT-associated promyelocytic leukaemia (PML) bodies were visualized by a combination of immunofluorescence with an anti-PML antibody and FISH using Alexa-488-(TTAGGG)_n PNA probes as described previously²¹.

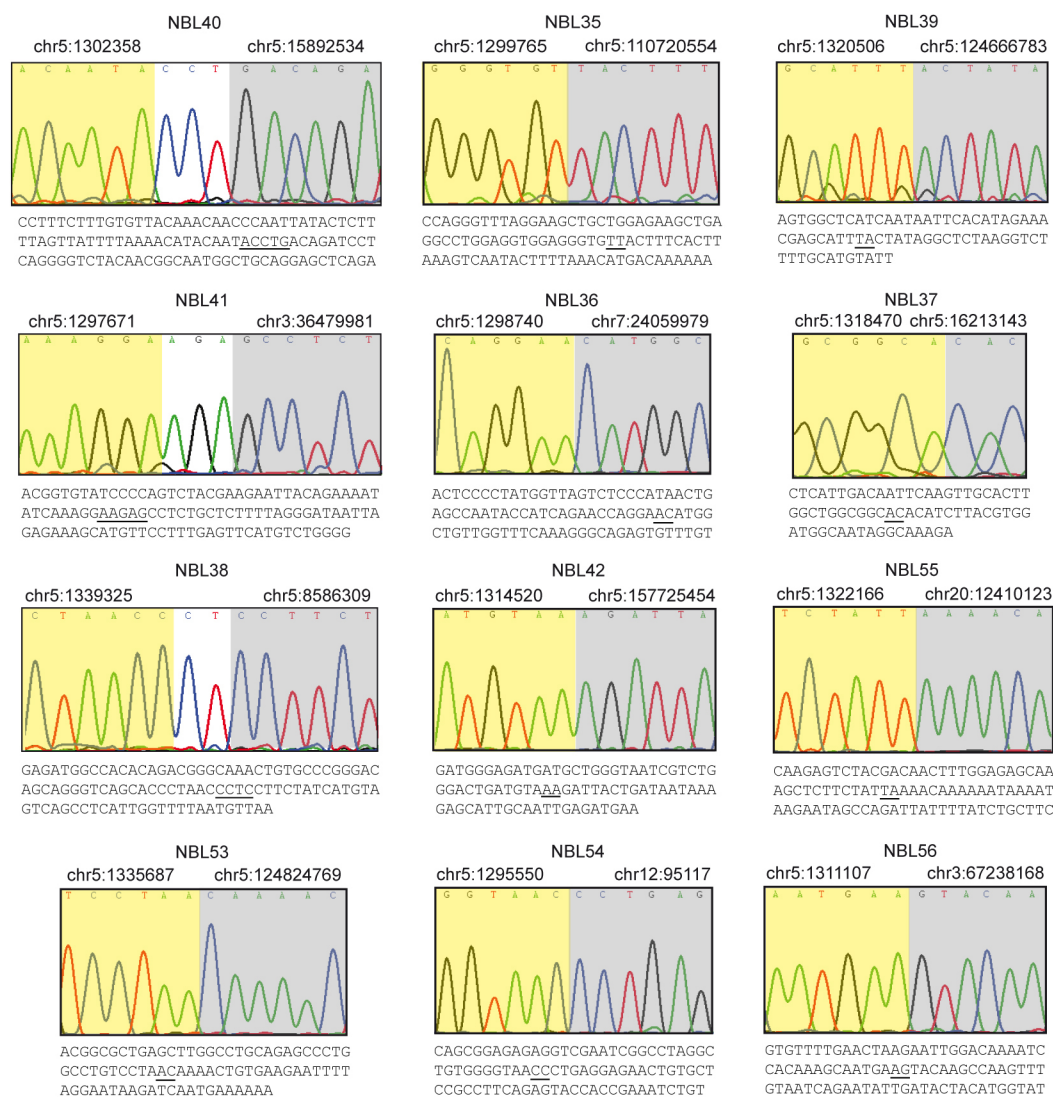
Telomere FISH on tumour sections. Tumours were sliced into 4 µm sections, paraffin fixed, embedded in formalin, and mounted onto positively charged glass microscope slides. Mounted sections were incubated for 30 min at 55 °C, washed three times for 5 min in xylene, rinsed in successive 100%, 95%, and 70% ethanol baths, and washed in double-distilled H₂O and 1% Tween before being placed in antigen unmasking solution in a boiling kitchen steam for 30 min. Next, slides were rinsed in double-distilled H₂O and dehydrated in successive ethanol washes of 70%, 95%, and 100%. Slides were incubated at 72 °C for 10 min with an Alexa-488 telomeric-C PNA probe and hybridized overnight in a dark humidity chamber. Slides were washed with PNA wash buffer and PBST and incubated for 10 min in DAPI solution. After washing in double-distilled H₂O, slides were mounted with prolong anti-fade mounting medium. Images were taken on a Nikon 90i fluorescent light microscope at ×63 resolution. Full z-stacks were taken at 0.5 µm and projected and focused using Elements software.

Statistical analyses. SPSS (package release 20.0.0, IBM Armonk), R (version 3.1.2), and GraphPad Prism (version 6.05 GraphPad Software) were applied for statistical analyses and data presentation. Overall survival was calculated as the time from diagnosis to death from disease or the last follow-up if the patient survived. Event-free survival was calculated from diagnosis to the time of tumour progression, relapse, or death from disease, or to the last follow-up if no event occurred. Survival curves were estimated according to Kaplan–Meier and compared with the log-rank test (R survival package version 2.15.0). Associations of genomic alterations with clinical risk factors were examined using Fisher's exact

test. Multivariable Cox regression models were used to analyse the simultaneous prognostic impact of *TERT* rearrangements and established clinical markers (stage (1–3, 4S versus 4), *MYCN* (non-amplified versus amplified), and age (<18 months versus >18 months)) on overall survival and event-free survival. Since *TERT* rearrangements were observed only in patients aged >18 months in this study, multivariable model building was restricted to this cohort ($n = 125$) and the variables *TERT* status, stage, and *MYCN* status. First, the proportional hazard assumption was assessed for each predictor one-at-a-time using the goodness-of-fit test of ref. 30 showing no deviation from the proportional hazard assumption. The proportional hazard assumption was considered valid whenever the P value of the goodness-of-fit test was >0.05. In addition, predicted survival curves under the Cox model were compared with the Kaplan–Meier estimates for each predictor supporting adequateness of model fit. Multivariable models were then built using a backwards selection procedure including the variables *TERT* status, stage, and *MYCN* status (inclusion criterion, P value of the score test ≤0.05; exclusion criterion, P value of the likelihood ratio test >0.1). The variables identified at this step formed the model of main effects. Finally, the factors selected in the model of main effects were fitted with all pairwise interactions in a second block by a stepwise forward selection (inclusion criterion, P value of the score test ≤0.05; exclusion criterion, P value of the likelihood ratio test >0.1), resulting in the final model. For the final model, the proportional hazard assumption was assessed using the goodness-of-fit test of ref. 30 as well as by fitting extended Cox models including the prognostic factors from the final model in a first block and the product terms of the prognostic factors with some function of time $g(t)$ in a second block with stepwise forward selection in the second block (inclusion criterion, P value of the score test ≤0.05; exclusion criterion, P value of the likelihood ratio test >0.1). Choices for $g(t)$ were $g(t) = t$ and $g(t) = \log(t)$ with t denoting survival time. The proportional hazard assumption was considered as valid if no time-dependent factor was selected in any of the extended Cox models and if, additionally, any P value of goodness-of-fit test was >0.05.

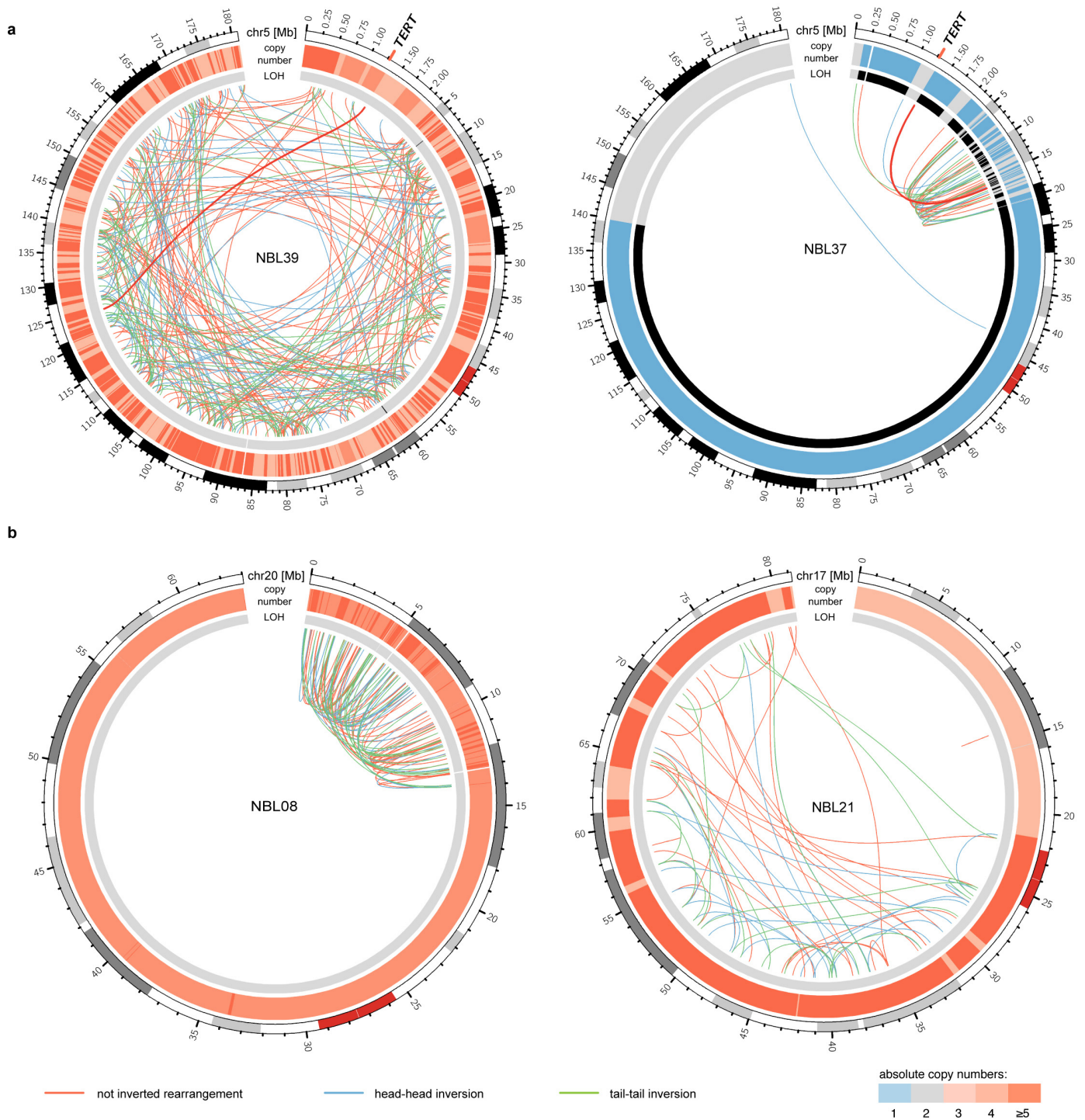
Analyses of *TERT* expression levels and methylation between subgroups were investigated by Mann–Whitney tests and corrected for multiple hypotheses testing using a Bonferroni correction. To test for mutual exclusivity between *TERT* rearrangements (*TERT*), *MYCN* amplifications (MNA), and *ATRX* mutations (*ATRX*) in the high-risk group, Fisher's exact tests were performed between every alteration and the combination of the remaining two alterations (*TERT* versus MNA and *ATRX*; MNA versus *TERT* and *ATRX*; *ATRX* versus *TERT* and MNA). The largest P value was finally reported.

25. Fernandez-Cuesta, L. *et al.* Frequent mutations in chromatin-remodelling genes in pulmonary carcinoids. *Nature Commun.* **5**, 3518 (2014).
26. Zhang, W. *et al.* Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. *Genome Biol.* **16**, 133 (2015).
27. Peng, X. *et al.* Tissue-specific transcriptome sequencing analysis expands the non-human primate reference transcriptome resource (NHPRT). *Nucleic Acids Res.* **43**, D737–D742 (2015).
28. Blecher-Gonen, R. *et al.* High-throughput chromatin immunoprecipitation for genome-wide mapping of *in vivo* protein-DNA interactions and epigenomic states. *Nature Protocols* **8**, 539–554 (2013).
29. Dahl, J. A. & Collas, P. A rapid micro chromatin immunoprecipitation assay (microChIP). *Nature Protocols* **3**, 1032–1045 (2008).
30. Harrell, F. & Lee, K. in *Proc. Eleventh Annual SAS User's Group International* 823–838 (SAS Institute, 1986).



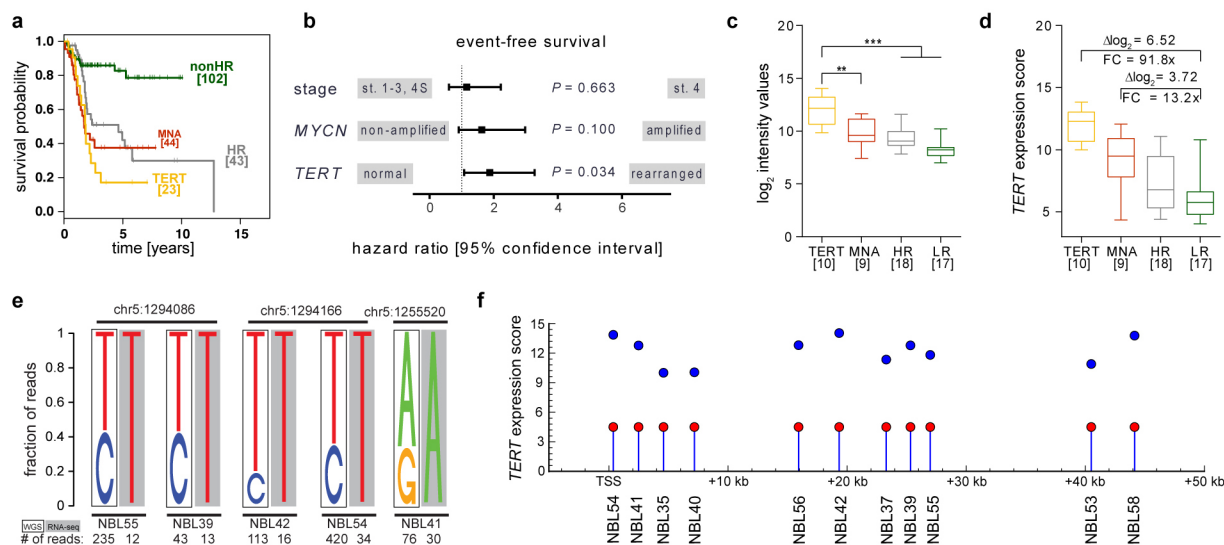
Extended Data Figure 1 | Validation of rearrangements of the *TERT* locus by dideoxy-sequencing. Sequencing chromatograms of the breakpoint regions of 5p15.33 rearrangements along with their genomic coordinates (hg19), and the breakpoint-spanning nucleotide sequences. The sequence

mapping to the *TERT* locus is indicated in yellow, the rearrangement partner is indicated in grey; nucleotides inserted at the breakpoint region are indicated in white.



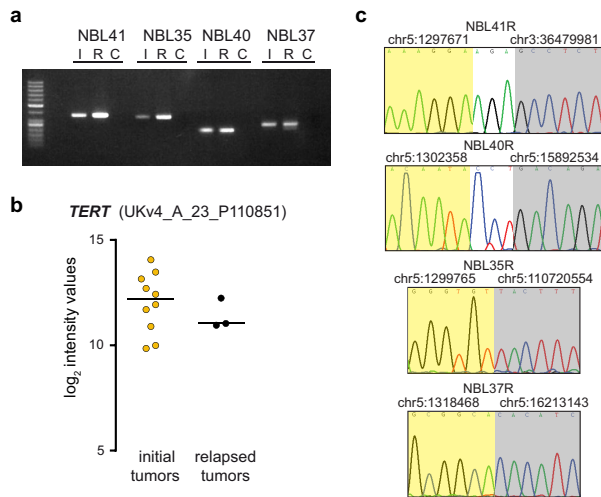
Extended Data Figure 2 | Schematic representation of chromothripsis in four primary neuroblastomas. a, Circos plots showing chromothripsis of chromosome 5 in the tumours NBL39 and NBL37. b, Circos plots

showing chromothripsis of chromosomes 17 and 20 in the tumours NBL21 and NBL08, respectively. Regions showing loss of heterozygosity (LOH) are indicated by black segments.

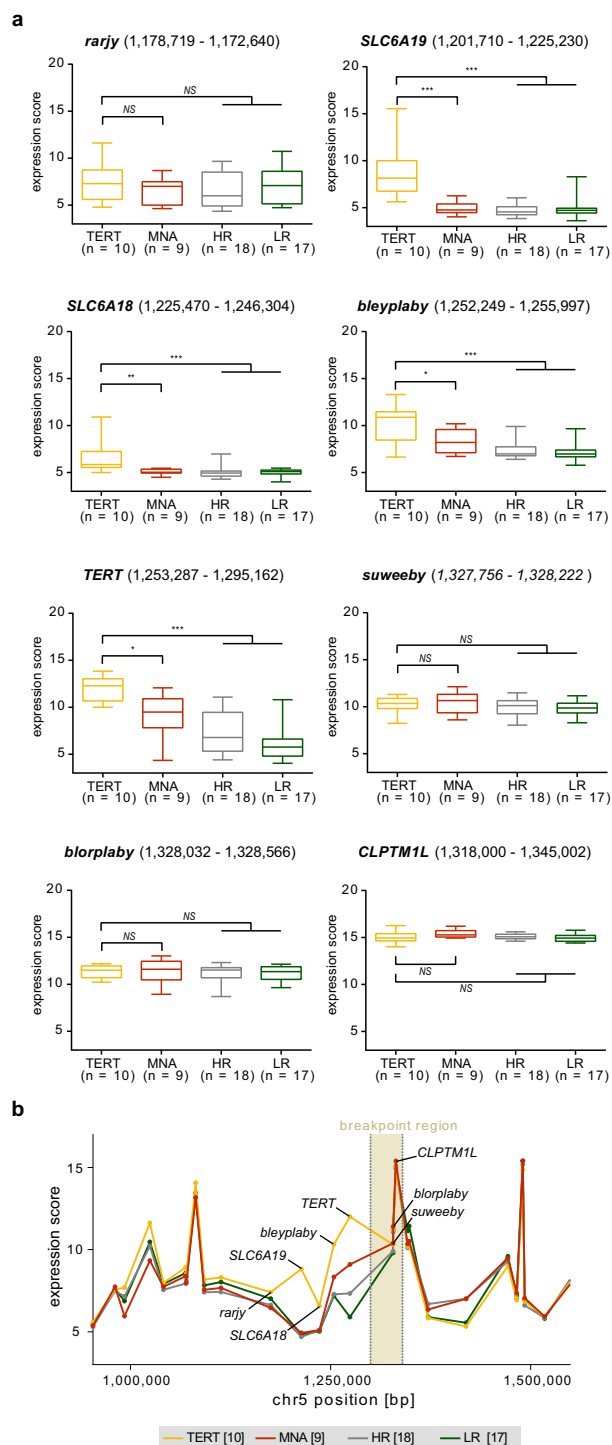


Extended Data Figure 3 | *TERT* rearrangements are associated with poor event-free survival of patients, and with *TERT* upregulation independent of the breakpoint distance from the *TERT* transcriptional start site. **a**, Kaplan-Meier estimates for event-free survival of neuroblastoma patient groups defined by *TERT* rearrangements (TERT), *MYCN* amplification (MNA), high-risk disease without these alterations (HR), and low-risk or intermediate-risk disease (nonHR). Patients with tumours bearing both a *TERT* rearrangement and *MYCN* amplification ($n = 5$) were excluded. Event-free survival at 5 years: 0.17 ± 0.09 (TERT) versus 0.38 ± 0.09 (MNA) versus 0.43 ± 0.09 (HR) versus 0.83 ± 0.05 (nonHR). **b**, Multivariable Cox regression analysis of the potential prognostic factors stage, *MYCN* status, and *TERT* status for event-free survival in patients aged >18 months ($n = 125$). **c**, Validation of *TERT* expression levels by microarrays in the four neuroblastoma subgroups indicated above. Sample numbers are given at the bottom. $^{**}P < 0.01$, $^{***}P < 0.001$. **d**, Fold-change of

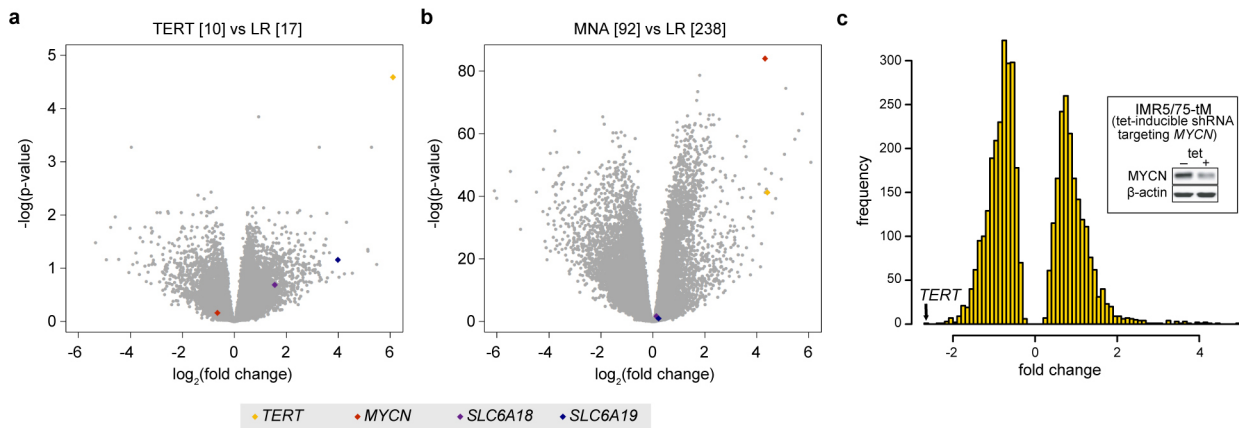
median *TERT* expression levels in *TERT*-rearranged and *MYCN*-amplified tumours compared with low-risk (LR) neuroblastomas as measured by transcriptome sequencing. **e**, Evidence for monoallelic expression of *TERT* in five tumours bearing *TERT* rearrangements. The presence of a heterozygous single nucleotide polymorphism and its allelic fraction measured by whole-genome sequencing (WGS) is shown on the left of each panel; monoallelic expression as established by transcriptome sequencing (RNA-seq) is indicated on the right. The genomic position of the single nucleotide polymorphism is indicated at the top, the number of reads available for the analysis is shown at the bottom. **f**, *TERT* expression measured by transcriptome sequencing in relation to the distance of the rearrangement breakpoint from the *TERT* transcriptional start site (TSS). *TERT* expression levels and breakpoint distances from the *TERT* transcriptional start site were not correlated ($r = 0$, $P = 0.97$; Spearman's rank correlation test).



Extended Data Figure 4 | *TERT* rearrangements are maintained in relapsed neuroblastoma. **a**, Agarose gel electrophoresis of PCR products representing individual *TERT* rearrangements in four tumours at initial diagnosis (I), and at relapse (R). The non-template controls are indicated by C. **b**, *TERT* expression measured by microarrays in relapsed *TERT*-rearranged tumours ($n = 3$) compared with *TERT*-rearranged tumours biopsied at initial diagnosis ($n = 10$). **c**, Sequencing chromatograms of the breakpoint regions for the relapse cases.

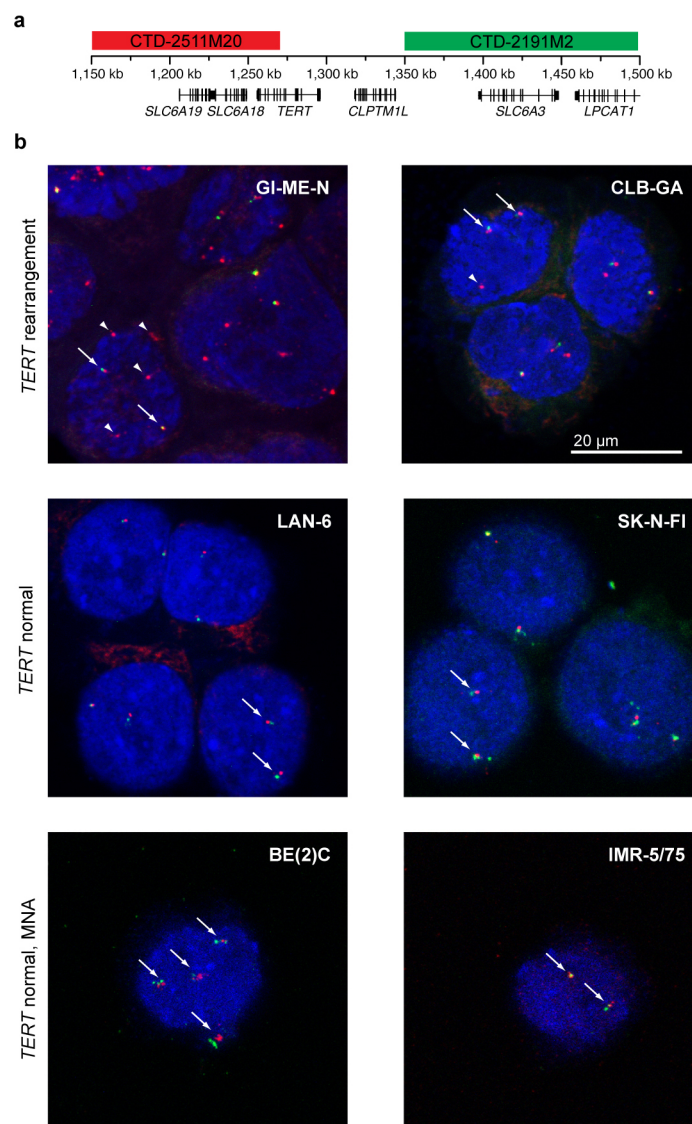


Extended Data Figure 5 | Regional effects of *TERT* rearrangements on gene expression patterns. **a**, Expression levels of genes around *TERT* measured by transcriptome sequencing are shown for tumours with *TERT* rearrangements (TERT, yellow), tumours with *MYCN* amplifications (MNA, red), and tumours without these aberrations classified as either high-risk (HR, grey) or low-risk (LR, green). Five consecutive genes (*TERT*, *bleyplaby*, *SLC6A18*, *SLC6A19*, and *rarjy*) located distal of the breakpoint at chromosome 5p15.33, and three genes (*suweeby*, *blorplaby*, and *CLPTM1L*) located proximal of the breakpoint are shown. Genomic positions of the genes are indicated at the top, and sample sizes are indicated at the bottom. NS, not significant, * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$. **b**, Averaged expression levels of genes measured by transcriptome sequencing are indicated for an ~500 kb region centring around *TERT*. Colour codes of neuroblastoma subgroups as indicated above. The breakpoint region is indicated in beige.



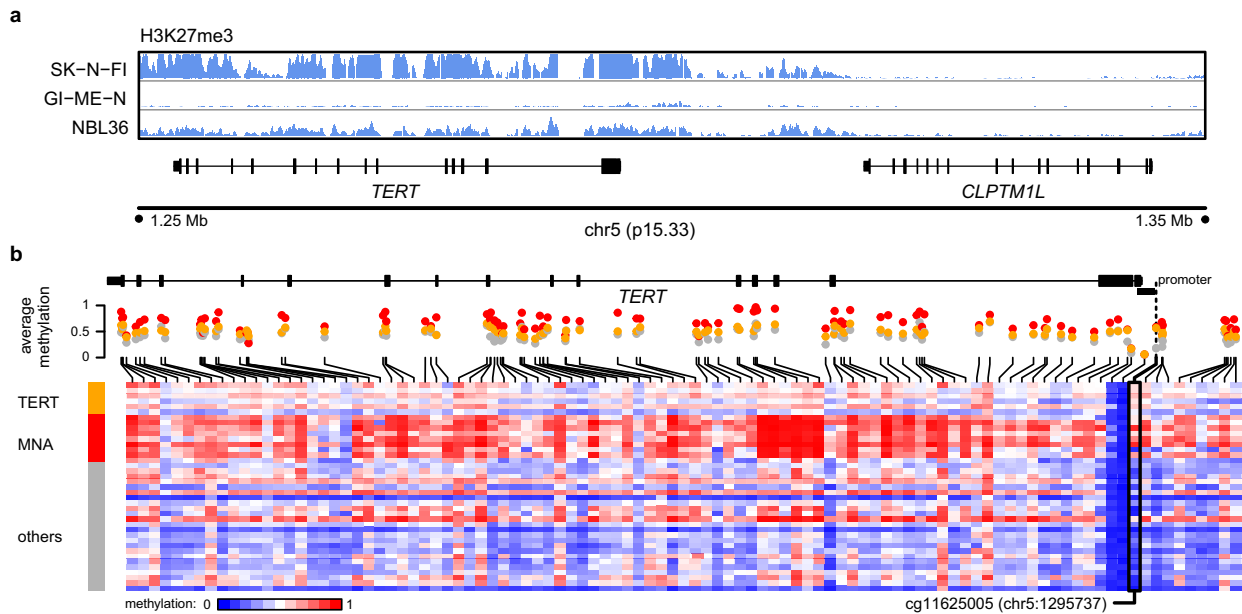
Extended Data Figure 6 | *TERT* mRNA levels are massively upregulated in *TERT*-rearranged and *MYCN*-amplified neuroblastomas. **a**, Differential gene expression between *TERT*-rearranged ($n = 10$) and low-risk ($n = 17$) neuroblastomas and **b**, between *MYCN*-amplified (MNA, $n = 92$) and low-risk tumours ($n = 238$) measured by transcriptome sequencing²⁶. Negative $\log(P)$ values are plotted against $\log_2(\text{fold changes})$ of the genes expressed in the respective subgroups. Genes of interest are indicated at the bottom. $\log_2(\text{fold changes})$ correspond to the log difference of established gene expression scores. P values were established using Student's t -tests and applying a Benjamini and Hochberg false discovery rate multiple testing correction.

c, Analysis of differential gene expression of a *MYCN*-amplified neuroblastoma cell line (IMR-5/75) before and after shRNA-mediated knockdown of *MYCN* using transcriptome sequencing. The number of genes (frequency) up- or downregulated upon *MYCN* knockdown is plotted against the fold change of regulation. *TERT* is the most strongly downregulated gene upon shRNA-mediated *MYCN* knockdown. Only differentially expressed genes (t -test, false discovery rate controlled with $P < 0.01$) are shown. The insert shows reduced *MYCN* protein levels after induction of a *MYCN*-specific shRNA in IMR-5/75 cells as well as β -actin protein levels as control analysed by immunoblotting.



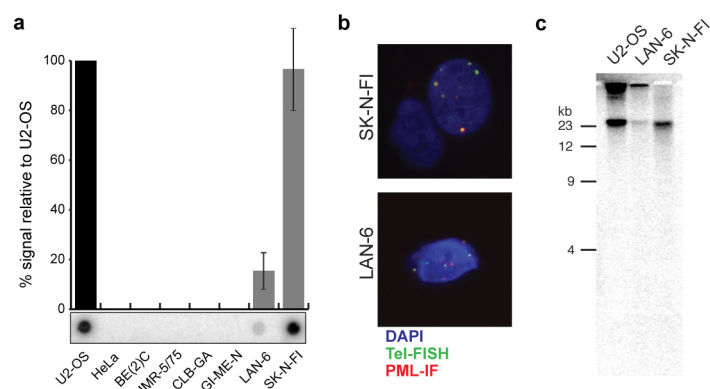
Extended Data Figure 7 | Detection of *TERT* rearrangements in neuroblastoma cell lines. **a**, Schematic representation of binding sites of probes used for FISH at the *TERT* locus. **b**, FISH analysis of neuroblastoma cell lines with 5p15.33 rearrangements and without these rearrangements; MNA,

MYCN-amplified. Arrows indicate the unaltered status of chromosome 5p15.33 (red signals in close proximity to green signals); arrowheads indicate a rearrangement of 5p15.33 (red signals without adjacent green signals).



Extended Data Figure 8 | Patterns of H3K27me3 histone modification and DNA methylation at the *TERT* locus. **a**, Survey of the histone mark H3K27me3 in neuroblastomas harbouring *TERT* rearrangements (cell line GI-ME-N and primary tumour NBL36) and a neuroblastoma cell line lacking these alterations (SK-N-FI). **b**, DNA methylation patterns of CpG sites at the *TERT* locus in *TERT*-rearranged (*TERT*, $n = 6$) and *MYCN*-amplified (MNA, $n = 9$) primary neuroblastomas, as well as tumours lacking these alterations (others, $n = 24$) using HumanMethylation450 microarrays.

Samples are ordered from top to bottom. CpG sites are indicated relative to their position to the *TERT* locus. Average methylation levels of each CpG site in the three subgroups are shown at the top. Similar to the highlighted CpG site cg11625005 upstream of the *TERT* transcriptional start site (Fig. 3d), CpG sites scattered over the *TERT* gene body were methylated significantly higher in *TERT*-rearranged and *MYCN*-amplified cases than in tumours without these alterations ($P < 0.001$ each).



Extended Data Figure 9 | ALT activity in neuroblastoma cell lines lacking *TERT* rearrangements and *MYCN* amplifications. **a**, Detection of extrachromosomal telomeric repeat DNA by C-circle assays with genomic *AluI*/*MboI* digested DNA from the indicated neuroblastoma cell lines. The ALT-positive control cell line, U2OS, is at the left. **b**, Detection of ALT-associated PML bodies in the indicated cell lines. Telomeric TTAGGG FISH

(green) and immunofluorescence for PML (red) were combined and DNA was counterstained with DAPI (blue). **c**, Telomere restriction-fragment analysis of telomeric DNA from LAN-6, SK-N-FI, and the ALT-positive control U2OS. Telomeric DNA was detected by Southern blot with a [32 P]dATP end-labelled (CCCTAA) $_4$ oligonucleotide.

Extended Data Table 1 | Prognostic characteristics of the neuroblastoma cohort and prevalence of *TERT* rearrangements in patient subgroups

a

Prognostic characteristics of the neuroblastoma cohort

	n	% of total
INSS stage		
1	31	14.3 %
2	28	12.9 %
3	25	11.5 %
4	104	47.9 %
4S	29	13.4 %
Age at diagnosis		
< 18 months	92	42.4 %
> 18 months	125	57.6 %
MYCN status		
Normal	168	77.4 %
Amplified (MNA)	49	22.6 %
Risk groups		
High-risk [HR] patients	114	52.5 %
- HR w/o MNA	65	30.0 %
Non-HR patients	103	47.5 %

b

Prevalence of *TERT* rearrangements in prognostic neuroblastoma subgroups

	n ALL	n <i>TERT</i> _r	odds ratio	P-value
INSS stage				
1-3, 4S	113	1		
4	104	27	38.81	<0.001
Age at diagnosis				
< 18 months	92	0		
> 18 months	125	28	∞	<0.001
Risk groups				
Non-HR patients	103	1		
High-risk [HR] patients	114	27	31.33	<0.001
MYCN status				
Normal	168	23		
Amplified (MNA)	49	5	[0.72]	0.633
HR w/o MNA	65	22		
HR w/ MNA	49	5	0.22	0.004

Thalamic control of sensory selection in divided attention

Ralf D. Wimmer^{1*}, L. Ian Schmitt^{1*}, Thomas J. Davidson², Miho Nakajima¹, Karl Deisseroth^{2,3,4} & Michael M. Halassa^{1,5,6}

How the brain selects appropriate sensory inputs and suppresses distractors is unknown. Given the well-established role of the prefrontal cortex (PFC) in executive function¹, its interactions with sensory cortical areas during attention have been hypothesized to control sensory selection^{2–5}. To test this idea and, more generally, dissect the circuits underlying sensory selection, we developed a cross-modal divided-attention task in mice that allowed genetic access to this cognitive process. By optogenetically perturbing PFC function in a temporally precise window, the ability of mice to select appropriately between conflicting visual and auditory stimuli was diminished. Equivalent sensory thalamocortical manipulations showed that behaviour was causally dependent on PFC interactions with the sensory thalamus, not sensory cortex. Consistent with this notion, we found neurons of the visual thalamic reticular nucleus (visTRN) to exhibit PFC-dependent changes in firing rate predictive of the modality selected. visTRN activity was causal to performance as confirmed by bidirectional optogenetic manipulations of this subnetwork. Using a combination of electrophysiology and intracellular chloride photometry, we demonstrated that visTRN dynamically controls visual thalamic gain through feedforward inhibition. Our experiments introduce a new subcortical model of sensory selection, in which the PFC biases thalamic reticular subnetworks to control thalamic sensory gain, selecting appropriate inputs for further processing.

To dissect the circuit basis of sensory selection, we sought a behaviour capable of dividing attention across modalities in the freely behaving mouse. Building on the rich history of visual neuroscience^{6–8}, we focused our investigations on visual processing under conditions in which vision was behaviourally selected or suppressed (Fig. 1a). As such, we developed and validated a two-alternative forced-choice task in which mice selected between conflicting visual and auditory stimuli on a trial-by-trial basis. Stimuli indicated the location where a mouse had to nose poke to obtain a reward. Trial availability and target modality were signalled through binaurally emitted noise. For some mice, brown noise (10-kHz low-pass-filtered white noise) signalled a visual target (Fig. 1b, top) whereas blue noise (11-kHz high-pass-filtered white noise) signalled an auditory target (Fig. 1b, bottom). Cueing was counterbalanced across mice, with no effect on performance noted. By design, the task was asymmetric, with a visual detection component (flash from a light-emitting diode (LED) appearing to the right or left) and an auditory discrimination component (upsweep, 10–14 kHz; downsweep, 16–12 kHz). Multiple quality control metrics ensured that mice performed this task using the biasing cues (brown and blue noise) rather than low-level alternating strategies (Extended Data Fig. 1). Performance on the two modalities was balanced ($n = 15$ mice, Fig. 1c).

Comparing visual detection under cross-modal and visual-only conditions suggested divided attention between vision and audition in the cross-modal task (Fig. 1d). Specifically, the visual detection threshold

was higher in cross-modal trials ($n = 4$ mice, $P < 0.05$, bootstrap comparison; see Extended Data Fig. 2 for single-mouse examples and fixed lapse rate analysis). This difference persisted even when the conflicting auditory stimulus was randomly but systematically removed (Fig. 1e and Extended Data Fig. 3), suggesting that diminished visual perception under cross-modal conditions is a result of expectation (top-down)

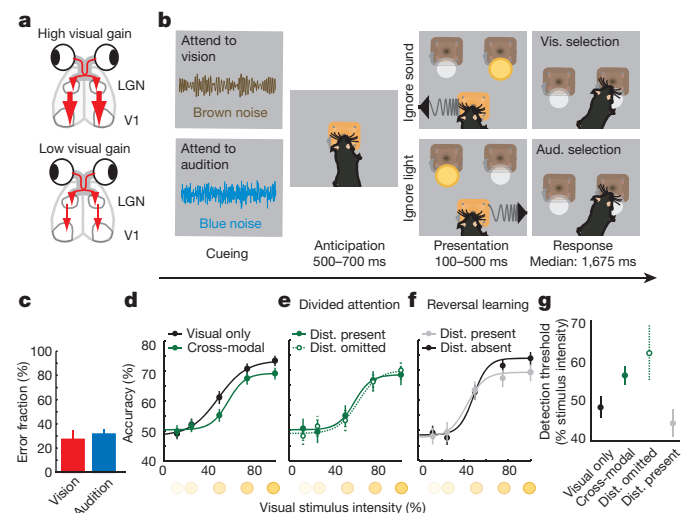


Figure 1 | Cross-modal divided attention in the mouse. **a**, Hypothesized control of visual gain under cross-modal conditions (LGN, lateral geniculate nucleus; V1, primary visual cortex). **b**, Task design. A mouse is simultaneously informed about trial availability and the nature of the target stimulus through binaurally delivered noise. In this schematic, brown noise denotes ‘attend to vision’ and blue noise denotes ‘attend to audition’. Following a variable anticipation, during which the mouse is required to hold its snout in a centrally located poke, conflicting auditory and visual stimuli are presented. By design, the task is asymmetric, having a visual detection component (presence or absence of light at the reward location) and an auditory discrimination component (upsweep, turn left; downsweep, turn right). **c**, Mice exhibited comparable performance on visual and auditory trials (mean \pm s.e.m., $n = 15$ mice). **d**, Visual detection performance in cross-modal trials compared to visual-only trials ($n = 4$ mice, ≥ 421 trials per condition). Note that both detection threshold and peak performance were lower in the cross-modal condition. **e**, Eliminating the auditory distractor in the cross-modal condition did not affect the visual detection psychometric function ($n = 4$ mice, ≥ 211 trials per condition). **f**, When mice were not differentially cued but instead ignored the auditory stimulus by learning that it was not rewarded over a full session (reversal learning), visual detection threshold did not change ($n = 6$ mice, ≥ 242 trials per condition). **g**, Visual detection threshold (bootstrap computed) of the pertinent psychometric functions in **d–f**. Error bars in **d–g** are 95% confidence intervals and therefore non-overlap denotes significance of $P < 0.05$. Aud., auditory; dist., distractor; vis., visual.

¹New York University Neuroscience Institute, Department of Neuroscience and Physiology, New York University Langone Medical Center, New York, New York 10016, USA. ²Department of Bioengineering, Stanford University, Stanford, California 94305, USA. ³Cracking the Neural Code Program, Stanford University, Stanford, California 94305, USA. ⁴Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, California 94305, USA. ⁵Department of Psychiatry, New York University Langone Medical Center, New York, New York 10016, USA. ⁶Center for Neural Science, New York University, New York, New York 10003, USA.

*These authors contributed equally to this work.

rather than sensory interference (bottom-up). Conversely, when mice selected targets based on reward history, detection threshold was unaffected (Fig. 1f, $n = 6$ mice). Selective suppression of the visual detection threshold in cross-modal trials was observed in both the raw (Fig. 1d, e) and the fitted (Fig. 1g) data. Together, these findings suggest that a cued, trial-by-trial task design is required for investigating sensory selection in divided attention.

Given the known role of PFC in top-down control of sensory processing and that our psychophysical measurements revealed top-down engagement in the cross-modal task, we asked whether cross-modal performance was PFC dependent. We targeted the prelimbic cortex because of its known homology to primate dorsolateral PFC⁹. We used the VGAT-ChR2 mouse to perturb PFC function in a temporally precise manner. In this mouse, the light-activated ion channel channelrhodopsin-2 (ChR2) is expressed under the vesicular GABA (γ -aminobutyric acid) transporter promoter (VGAT)¹⁰. Optogenetic drive in this mouse is known to exert intensity- and duration-dependent inhibition of excitatory neural activity^{11,12}. Using this approach, we observed behavioural disruption only when the PFC activity was perturbed during stimulus anticipation (Fig. 2a, $n = 4$

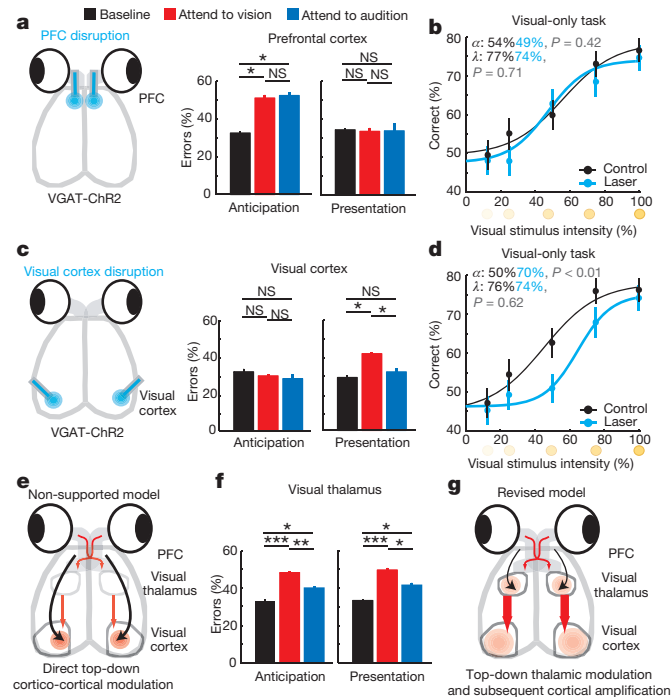


Figure 2 | Evidence for top-down thalamic modulation in divided attention.

a, Disrupting PFC activity by delivering blue laser pulses (50 Hz, 18 ms, 90% duty cycle) impaired task performance at 100% stimulus intensity equally on both modalities only when manipulation was performed during stimulus anticipation ($n = 4$ mice, $*P < 0.05$, Wilcoxon rank-sum test). **b**, The effect was related to the cross-modal nature of the task, not its difficulty, as PFC inhibition did not affect performance on a visual-only task. **c**, Disruption of primary visual cortex during stimulus presentation impaired performance on visual trials ($n = 4$ mice). **d**, The effect in **c** was related to task difficulty, as the visual detection threshold increased in a visual-only task. **e**, The data in **a** and **c** do not support a causal role for PFC interactions with primary visual cortex in performance. **f**, Perturbing visual thalamic function in a manner similar to cortical perturbations in VGAT-ChR2 mice preferentially diminished performance on visual trials during both anticipation and presentation of target stimuli ($n = 12$ sessions from 3 mice, $*P < 0.05$, $***P < 0.001$, Wilcoxon rank-sum test). **g**, The finding in **f** supports a model in which PFC activity influences thalamic sensory processing. Bar graphs represent mean \pm s.e.m. Error bars for psychometric curves are 95% confidence intervals. NS, not significant; α , detection threshold; λ , lapse rate.

mice). This effect was dependent on the cross-modal nature of the task, as perturbing PFC activity in a visual-only task had no effect on performance regardless of task difficulty (Fig. 2b, $n = 3$ mice, ≥ 246 trials per condition).

We reasoned that the PFC might be exerting its effect on performance by biasing sensory circuits towards target stimulus processing and distractor suppression. Several studies have suggested that this top-down bias is exerted at the level of the sensory cortex^{4,13}. We did not find this to be the case in our task; perturbing visual cortical activity diminished visual performance only during stimulus presentation (Fig. 2c). This effect, unlike that of the PFC, was not cross-modal task specific, as it increased the detection threshold in a visual-only task (Fig. 2d, $n = 3$ mice, ≥ 239 trials per condition, $P < 0.01$, bootstrap comparison). An analogous manipulation of auditory cortex resulted in a qualitatively similar effect on performance, but with a larger effect size (Extended Data Fig. 4a, $n = 4$ mice, $***P < 0.01$, Wilcoxon rank-sum test), probably owing to the auditory cortical requirement for stimulus discrimination¹⁴. Together, these findings support the role of sensory cortical areas in stimulus amplification and discrimination, but are inconsistent with their being a locus for top-down bias of sensory processing. Moreover, optogenetic perturbations of frontal regions that project directly to visual and auditory cortices such as the anterior cingulate cortex (ACC) and the lateral orbitofrontal cortex (OFC) did not affect cross-modal performance (Extended Data Fig. 4b, c, $n = 8$ sessions, 2 mice). In contrast, localized viral injection of AAV-hSyn-DIO-ChR2 into the prelimbic cortex of mice in which Cre recombinase is expressed in inhibitory neurons (VGAT-Cre), followed by optical manipulation, disrupted performance (Extended Data Fig. 4d–h).

Having shown that direct prefrontal–sensory cortical interaction is unlikely to account for top-down control of visual processing (Fig. 2e), we sought to find the locus of attentional modulation observed in the cross-modal task (Fig. 1c). The sensory thalamus has been implicated in attentional modulation in primates^{15–17}, raising the possibility that it could be a locus of top-down sensory bias. Using a similar VGAT-ChR2 strategy, we found that optogenetic perturbation of the visual thalamus (lateral geniculate nucleus; LGN) during either stimulus anticipation or presentation impaired cross-modal performance (Fig. 2f, $n = 12$ sessions, 3 mice). This suggested that top-down bias of visual detection may be based in the thalamus (Fig. 2g). Visual thalamic manipulation resulted in worsening performance on both auditory and visual trials, implicating intrathalamic interactions in this behavioural effect. Intrathalamic interactions are mediated by a shell of GABAergic neurons surrounding thalamic nuclei known as the thalamic reticular nucleus (TRN)¹⁸. TRN neurons have been hypothesized to control the gain of thalamic output in a behaviourally relevant manner^{19,20}. As such, we asked whether the TRN could be a locus of top-down modulation of the sensory thalamus.

To investigate the role of the TRN in visual gain control in divided attention, we used an intersectional genetic/connectivity strategy to label inhibitory (VGAT-positive) neurons that project to the LGN with retrograde lentiviruses²⁰ (Fig. 3a and Extended Data Fig. 5a, b) and optogenetically tagged them during extracellular recordings in freely behaving mice (Extended Data Fig. 5c–f).

During cross-modal performance, we observed bidirectional modulation of visTRN neurons consistent with their hypothesized role in behaviourally relevant sensory gain control. Specifically, during ‘attend to vision’ trials, firing rates of these neurons were reduced. In contrast, their rates were elevated during ‘attend to audition’ trials (Fig. 3b (example), Fig. 3c (population; $n = 138$ cells) and Extended Data Fig. 6). Attentional modulation by the visTRN was eliminated by optogenetic PFC disruption (Fig. 3d, e, $n = 56$ cells). Although PFC disruption diminished performance, the effects on visTRN firing rates were not simply a covariant of inaccurate performance (Extended Data Fig. 7). In contrast to naturally occurring error trials, during which

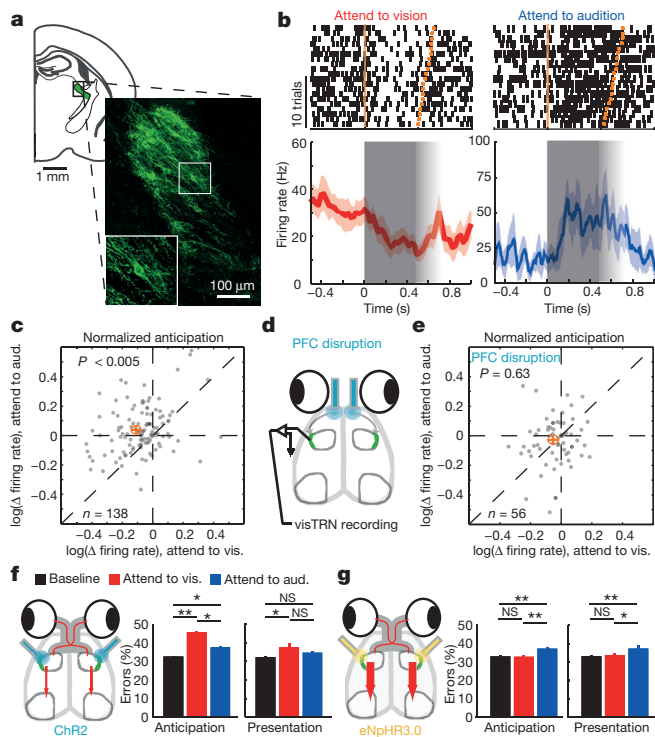


Figure 3 | PFC-dependent visTRN modulation suggests PFC-TRN functional coupling is required for visual gain control. **a**, Intersectional tagging of visTRN neurons based on connectivity and genetic identity. Inset, maximum projection of ten 1- μ m confocal images. Cells were labelled with Chr2-eYFP and stained with anti-GFP. eYFP, enhanced yellow fluorescent protein; GFP, green fluorescent protein. **b**, Raster of two visTRN neurons triggered on task initiation. Note the reduction of firing rate between the first trigger (trial initiation) and the second one (stimulus presentation) during the 'attend to vision' condition, but the opposite during the 'attend to audition' condition. Red dotted lines indicate triggers. Fading grey boxes denote the jitter of the anticipatory period. **c**, Group analysis of **b**, showing a scatter plot of responses from visTRN neurons ($n = 138$ cells from 4 animals, $P < 0.005$, Wilcoxon rank-sum test performed over all cells). The orange crosshair indicates mean \pm 95% confidence interval. **d**, PFC activity was disrupted during stimulus anticipation to examine the effect on visTRN activity. **e**, PFC disruption diminished visTRN modulation of attention. **f, g**, Behavioural performance is causally dependent on visTRN modulation of attention. **f**, Optogenetic activation of retrograde-tagged visTRN neurons resulted in performance diminishing on visual trials (mean \pm s.e.m., $n = 12$ sessions from 3 mice, $*P < 0.05$, $**P < 0.01$, $***P < 0.001$, Wilcoxon rank-sum test), consistent with this manipulation lowering visual gain. **g**, In contrast, optogenetic inhibition of visTRN neurons preferentially diminished performance on auditory trials, consistent with inappropriate visual gain increase ($n = 12$ sessions from 3 mice).

diminished visTRN firing rate modulation was observed in both 'attend to vision' and 'attend to audition' trials, PFC disruption had a greater effect on 'attend to audition' trials. This result may highlight a role of the PFC in distractor suppression.

To examine whether the physiological effects observed in the visTRN were causal to behaviour, we manipulated this subnetwork bidirectionally. Although activating the visTRN resulted in similar effects to driving LGN inhibition in VGAT-ChR2 mice (Fig. 3f, $n = 12$ sessions, 3 mice), the effect size was smaller, probably reflecting the weaker nature of the genetic manipulation (Extended Data Fig. 8). This result supports the notion that elevated visTRN firing reduces visual thalamic gain. In contrast, inhibiting visTRN function using the light-activated Cl^- pump eNpHR3.0 diminished performance on 'attend to audition' trials, suggesting it inappropriately enhanced visual thalamic gain when it needed to be suppressed (Fig. 3g, $n = 12$ sessions, 3 mice).

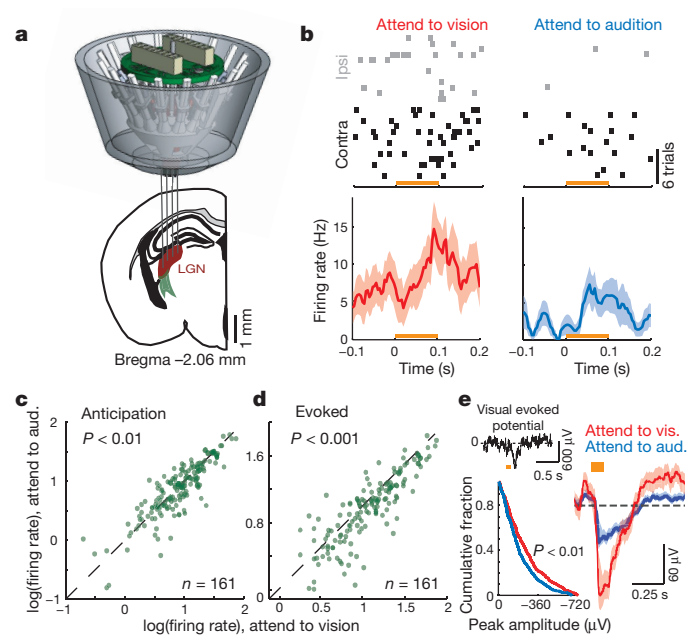


Figure 4 | Direct evidence for visual thalamic gain modulation in divided attention. **a**, Cartoon depiction of multi-electrode targeting of the LGN at 2.06 mm posterior to Bregma in freely behaving mice. **b**, An example of differential modulation of a single LGN cell spiking under the two anticipatory task conditions. Note that contralateral eye stimulation (with respect to recording electrodes) resulted in more robust visual drive. Moreover, the cell discharged more spikes during anticipation and presentation when attention was directed towards vision. Contra, contralateral; Ipsi, ipsilateral. **c, d**, Group analysis of the phenomenon in **b** ($n = 161$ cells from 4 mice, Wilcoxon rank-sum test). **e**, Similarly, enhanced visual responses were observed at the level of visual evoked potentials. Top left, example visual evoked potential (VEP); bottom left, cumulative distribution of VEP amplitudes, showing higher values for 'attend to vision' trials ($P < 0.01$, Kolmogorov-Smirnov test). Right, average VEP from 4 mice (684 visual trials and 633 auditory trials from 29 sessions; dashed line indicates baseline potential). Shaded errors are 95% confidence intervals. Orange blocks denote visual stimulus presentation.

To determine whether modulation of visTRN firing rate affected visual processing, we investigated LGN spiking in response to well-controlled visual stimuli. We implanted mice with multi-electrode arrays targeted to the LGN (Fig. 4a). To minimize trial-to-trial variability related to slight changes in head and eye position, we changed the position of visual stimuli from wall-mounted to head-mounted LEDs. LGN neurons showed enhanced baseline and evoked activity when attention was directed to vision (Fig. 4b (example), Fig. 4c, d, (population, $n = 161$ cells, 4 mice)), consistent with baseline and evoked neuroimaging results observed in human LGN²¹. Differences in evoked responses were also observed in the visual evoked potential (Fig. 4e, $n = 684$ visual trials and $n = 633$ auditory trials, 4 mice). These physiological effects were not observed during error trials (Extended Data Fig. 9).

Suppression of LGN spiking in 'attend to audition' trials could be a result of direct feedforward inhibition or reduction in feedforward excitation. In contrast to many extrareticular inhibitory inputs²², the visTRN is known to exert direct feedforward inhibition on LGN neurons²³ (Fig. 5a). As such, we sought to measure LGN inhibition directly. To do so, we leveraged a recently developed technique known as fibre photometry, which has been used to measure bulk changes of intracellular Ca^{2+} concentration ($[\text{Ca}^{2+}]_i$) (ref. 24). We modified two aspects of conventional photometry to allow interrogation of intracellular Cl^- concentration ($[\text{Cl}^-]_i$), a proxy for synaptic inhibition and an otherwise inaccessible measure. First was the genetic labelling of neurons with the Cl^- indicator SuperClomeleon²⁵. This fluorescence resonance energy transfer (FRET) indicator contains cyan fluorescent

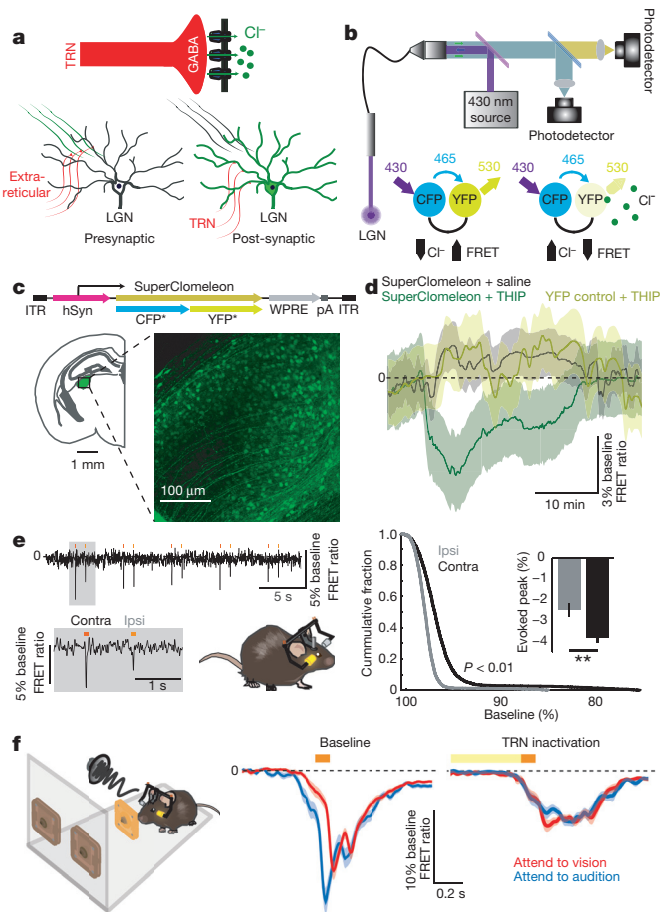


Figure 5 | Measuring bulk intracellular $[Cl^-]$ *in vivo* shows dynamic changes in LGN inhibition during behaviour. **a**, Possible mechanisms for modulation of LGN firing rate: extrareticular inputs can change activity by presynaptic inhibition of feedforward excitation, whereas visTRN inhibits LGN directly. **b**, FRET photometry setup and schematic of CFP-to-YFP FRET. **c**, Cloning of SuperCameleon, a FRET-based Cl^- indicator, into an adeno-associated virus (AAV) followed by *in vivo* expression in the LGN. hSyn, human synapsin promoter; ITR, inverted terminal repeat sequences; pA, polyadenylation signal; WPRE, woodchuck hepatitis virus post-transcriptional regulatory element. **d**, Pharmacological confirmation of the technical feasibility of SuperCameleon FRET for GABA_A-mediated increase in intracellular $[Cl^-]$ by injection of the GABA_A agonist THIP. Note that the YFP control mice did not show similar signals ($n = 3$ mice per condition; shaded errors are 95% confidence intervals). **e**, Mice showed stronger visual-evoked SuperCameleon FRET responses for the contralateral eye, as would be predicted ($n = 3$ mice, $P < 0.05$, Wilcoxon rank-sum test). Yellow bars mark the display of the light stimuli. **f**, Left, cartoon depiction of photometry in the cross-modal task, during which the visual stimulus was signalled through a head-mounted LED as in Fig. 4 (see Supplementary Video 1 for illustration); middle, differential visual-evoked $[Cl^-]_i$ LGN responses in relation to the modality anticipated (363 correct visual trials and 274 correct auditory trials from 6 mice). Shaded errors are 95% confidence intervals. Note that 'attend to audition' trials showed an earlier increase in $[Cl^-]_i$ (decreased SuperCameleon FRET) and the separation between the two traces started before stimulus onset, consistent with differential anticipatory changes of visTRN activity; right, optogenetic TRN inactivation eliminates this differential response (101 correct visual and 82 correct auditory trials from 3 mice). Orange bars indicate visual stimulus presentation and the yellow bar denotes optogenetic TRN inactivation.

protein (CFP) as FRET donor and yellow fluorescent protein (YFP) as FRET acceptor. Under conditions of elevated $[Cl^-]_i$, YFP is quenched and FRET is reduced (Fig. 5b). Second was the light path; we excited SuperCameleon with 430-nm light and collected both CFP and YFP emission data to perform subsequent ratiometric measurements off-line (Fig. 5b). To use this technology *in vivo*, we generated a viral

construct harbouring SuperCameleon (AAV-hSyn-SuperCameleon) and injected it into the LGN (Fig. 5c).

We validated SuperCameleon FRET as a measure of inhibition by two methods. First, we reasoned that pharmacological activation of GABA_A receptors would increase $[Cl^-]_i$ and reduce YFP emission. Indeed, intraperitoneal injection of the GABA_A agonist 4,5,6,7-tetrahydroisoxazolo(5,4-c)pyridin-3-ol (THIP, 8 mg kg⁻¹) reduced SuperCameleon FRET in a pharmacokinetically predicted manner²⁶ (Fig. 5d, $n = 3$ mice). Second, we expected visual drive to result in inhibition of LGN neurons that balanced excitation. This was indeed the case; these signals were larger when visual stimuli were delivered to the side contralateral to the recorded LGN than when delivered to the ipsilateral side (Fig. 5e, $n = 3$ mice). These rapid events were not observed in YFP control mice (Extended Data Fig. 10a). SuperCameleon visual transients were sensitive to the GABA_A receptor antagonist flumazenil in a dose-dependent manner (Extended Data Fig. 10b), confirming that these signals reflect GABAergic inhibition.

Having validated chloride photometry, we asked whether changes in visual gain associated with sensory selection were explained by opposing changes in LGN feedforward inhibition. We found this to be the case; visual-evoked chloride photometry showed significantly larger responses in 'attend to audition' trials than in 'attend to vision' trials (Fig. 5f, $n = 363$ visual trials and $n = 274$ auditory trials, 6 mice; example trials shown in Supplementary Video 1). Signal kinetics in these two conditions were distinct: an earlier reduction in FRET ratio was seen in auditory selection trials compared to visual selection trials, consistent with the differential baseline spiking observed in visTRN neurons. Optogenetic visTRN inactivation eliminated this differential inhibitory response (Fig. 5f and Extended Data Fig. 10c, $n = 3$ mice, >82 trials). Overall, our data support the model that thalamic gain control can be explained by feedforward inhibition and that the TRN is the source of this inhibition. More generally, to our knowledge, this experiment constitutes the first measurement of inhibitory dynamics in freely behaving animals.

Seminal studies have shown the thalamus to be more than a cortical relay^{27,28}. By providing a mechanistic circuit dissection of thalamic involvement in divided attention, we extend these studies in two directions. First, our findings in mice show the generality of thalamic modulation of attention across mammalian brains. Second, we provide a first description, with causal circuit dependence, of how prefrontal top-down control changes thalamic inhibitory dynamics to modulate sensory gain. The specific involvement of prelimbic cortex in this behaviour, which we further demonstrate through combined optogenetics and chloride photometry (Extended Data Fig. 10d), does not eliminate the possibility that the OFC and the ACC may be engaged in other types of top-down control, potentially via cortico-cortical interactions⁵. In addition to regulating sensory gain, prelimbic control of thalamic inhibition may regulate the degree by which relay nuclei participate in large-scale functional interactions¹⁷.

The ability to directly measure $[Cl^-]_i$ dynamics provided access to a critical biological variable: GABA_A-mediated synaptic inhibition. Although photometry has already been introduced into neuroscience for measurement of $[Ca^{2+}]_i$ in cell bodies and terminals²⁵, FRET-based chloride photometry has not been performed previously. In this study, developing chloride photometry was essential for establishing a direct physiological link between visTRN and LGN spiking (technical discussion in Supplementary Information).

Thirty years ago, Francis Crick proposed that the TRN functions as a 'searchlight', directing the internal spotlight of attention to thalamo-cortical circuits that process ongoing behavioural demands¹⁹. Owing to technical limitations, this transformative model has been difficult to test, particularly under conditions where the attentional spotlight shifts. Our study combined novel and established technology to provide mechanistic details for Crick's 'searchlight hypothesis', thereby contributing to understanding the circuit mechanisms of sensory selection.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 12 April; accepted 18 August 2015.

Published online 21 October 2015.

1. Miller, E. K. & Cohen, J. D. An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* **24**, 167–202 (2001).
2. Buschman, T. J. & Miller, E. K. Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. *Science* **315**, 1860–1862 (2007).
3. Fritz, J., Shamma, S., Elhilali, M. & Klein, D. Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nature Neurosci.* **6**, 1216–1223 (2003).
4. Rodgers, C. C. & DeWeese, M. R. Neural correlates of task switching in prefrontal cortex and primary auditory cortex in a novel stimulus selection task for rodents. *Neuron* **82**, 1157–1170 (2014).
5. Zhang, S. *et al.* Selective attention. Long-range and local circuits for top-down modulation of visual cortex processing. *Science* **345**, 660–665 (2014).
6. Glickfeld, L. L., Histed, M. H. & Maunsell, J. H. Mouse primary visual cortex is used to detect both orientation and contrast changes. *J. Neurosci.* **33**, 19416–19422 (2013).
7. Hubel, D. H. & Wiesel, T. N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol. (Lond.)* **160**, 106–154 (1962).
8. Newsome, W. T., Britten, K. H. & Movshon, J. A. Neuronal correlates of a perceptual decision. *Nature* **341**, 52–54 (1989).
9. Hoover, W. B. & Vertes, R. P. Anatomical analysis of afferent projections to the medial prefrontal cortex in the rat. *Brain Struct. Funct.* **212**, 149–179 (2007).
10. Vong, L. *et al.* Leptin action on GABAergic neurons prevents obesity and reduces inhibitory tone to POMC neurons. *Neuron* **71**, 142–154 (2011).
11. Halassa, M. M. *et al.* Selective optical drive of thalamic reticular nucleus generates thalamic bursts and cortical spindles. *Nature Neurosci.* **14**, 1118–1120 (2011).
12. Zhao, S. *et al.* Cell type-specific channelrhodopsin-2 transgenic mice for optogenetic dissection of neural circuitry function. *Nature Methods* **8**, 745–752 (2011).
13. Fritz, J. B., David, S. V., Radtke-Schuller, S., Yin, P. & Shamma, S. A. Adaptive, behaviorally gated, persistent encoding of task-relevant auditory information in ferret frontal cortex. *Nature Neurosci.* **13**, 1011–1019 (2010).
14. Letzkus, J. J. *et al.* A disinhibitory microcircuit for associative fear learning in the auditory cortex. *Nature* **480**, 331–335 (2011).
15. McAlonan, K., Cavanaugh, J. & Wurtz, R. H. Guarding the gateway to cortex with attention in visual thalamus. *Nature* **456**, 391–394 (2008).
16. Purushothaman, G., Marion, R., Li, K. & Casagrande, V. A. Gating and control of primary visual cortex by pulvinar. *Nature Neurosci.* **15**, 905–912 (2012).
17. Saalmann, Y. B., Pinsk, M. A., Wang, L., Li, X. & Kastner, S. The pulvinar regulates information transmission between cortical areas based on attention demands. *Science* **337**, 753–756 (2012).
18. Pinault, D. The thalamic reticular nucleus: structure, function and concept. *Brain Res. Brain Res. Rev.* **46**, 1–31 (2004).
19. Crick, F. Function of the thalamic reticular complex: the searchlight hypothesis. *Proc. Natl Acad. Sci. USA* **81**, 4586–4590 (1984).
20. Halassa, M. M. *et al.* State-dependent architecture of thalamic reticular subnetworks. *Cell* **158**, 808–821 (2014).
21. O'Connor, D. H., Fukui, M. M., Pinsk, M. A. & Kastner, S. Attention modulates responses in the human lateral geniculate nucleus. *Nature Neurosci.* **5**, 1203–1209 (2002).
22. Chen, C. & Regehr, W. G. Presynaptic modulation of the retinogeniculate synapse. *J. Neurosci.* **23**, 3130–3135 (2003).
23. Cox, C. L., Huguenard, J. R. & Prince, D. A. Nucleus reticularis neurons mediate diverse inhibitory effects in thalamus. *Proc. Natl Acad. Sci. USA* **94**, 8854–8859 (1997).
24. Gunaydin, L. A. *et al.* Natural neural projection dynamics underlying social behavior. *Cell* **157**, 1535–1551 (2014).
25. Grimley, J. S. *et al.* Visualization of synaptic inhibition with an optogenetic sensor developed by cell-free protein engineering automation. *J. Neurosci.* **33**, 16297–16309 (2013).
26. Cremers, T. & Ebert, B. Plasma and CNS concentrations of Gaboxadol in rats following subcutaneous administration. *Eur. J. Pharmacol.* **562**, 47–52 (2007).
27. Casagrande, V. A., Sáry, G., Royal, D. & Ruiz, O. On the impact of attention and motor planning on the lateral geniculate nucleus. *Prog. Brain Res.* **149**, 11–29 (2005).
28. Mitchell, A. S. *et al.* Advances in understanding mechanisms of thalamic relays in cognition and behavior. *J. Neurosci.* **34**, 15340–15346 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank J. A. Movshon, W. Ma, R. W. Tsien, G. Fishell and D. Rinberg for helpful comments on the manuscript and G. J. Augustine for providing us with the *SuperClomeleon* construct and for helpful discussion around its use. The work was supported by the Swiss National Science Foundation (P2LAP3 151786) to R.D.W. and the Simons Foundation, the Sloan Foundation, the Brain and Behavior Research Foundation and the US National Institutes of Health (R00 NS078115) to M.M.H.; M.M.H. is additionally supported by the Feldstein Medical Foundation, a Klingenstein-Simons Fellowship and a Biobehavioral Research Award for Innovative New Scientists (BRAINS) R01 (R01 MH107680) from the National Institute of Mental Health.

Author Contributions M.M.H. conceived and designed all aspects of the study. R.D.W. devised the training paradigm for the cross-modal task and L.I.S. performed all associated programming. R.D.W. collected electrophysiological data. T.J.D. provided fibre photometry training, advice and rig designs; L.I.S. extended the method to FRET-based photometry, built the rig and collected data. R.D.W. analysed behavioural data and L.I.S. analysed psychophysical, electrophysiological and photometry data. M.N. generated the retrograde lentiviruses in-house, performed *SuperClomeleon* cloning into an AAV backbone and acquired confocal images. K.D. provided support for fibre photometry training. M.M.H. supervised the experiment, directed the analysis and wrote the manuscript. All authors read the final version of the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.M.H. (michael.halassa@nyumc.org).

METHODS

Animals. VGAT-ChR2 mice were purchased from the Jackson Laboratory and maintained on a C57Bl6/J background. VGAT-Cre mice were backcrossed to C57Bl6/J mice for at least six generations. For experiments in Fig. 1, a total of fifteen animals were trained, ten of which were later used to establish psychometric functions (four for divided attention; six for reversal learning). For Fig. 2, four VGAT-ChR2 mice were used for disruption of PFC and primary sensory cortices and three mice were used for inactivating LGN. In Fig. 3, four VGAT-Cre mice were used for electrophysiological recordings from optogenetically identified visTRN neurons, of which two were used for combined electrophysiological recordings with optogenetic PFC inactivation (Fig. 3c–e). An additional six mice were used for optogenetic activation or inhibition of visTRN (three per manipulation) during behaviour (Fig. 3f–g). Four wild-type mice were used for LGN recordings (Fig. 4). For fibre photometry experiments (Fig. 5), six mice were injected with AAV-hSyn-SuperClomeleon for behavioural and pharmacological experiments and three YFP control mice were used. Including all animals used for Extended Data figures, a total of 28 male mice, 1.5–6 months old, were trained on the cross-modal task. All experimental procedures involving animals were performed according to the guidelines of the Institutional Animal Care and Use Committee at the New York University Langone Medical Center and the US National Institutes of Health.

Behavioural training and testing. *Behavioural setup.* Experiments were conducted in a custom-built trapezoidal testing chamber (base 1, 12 cm; base 2, 25 cm; height, 25 cm) positioned over a grid floor. The testing chamber contained three nose-pokes, each of which consisted of an infrared LED/infrared phototransistor pair (Digikey, Thief River Falls, Minnesota) for response detection. Activation of a central nose-poke located on the grid floor, 6 cm away from the reward wall, was required for trial initiation. Two headphone speakers (Skullcandy, Park City, Utah) embedded in the floor delivered biasing cues binaurally. Two white LEDs (Mouser, El Cajon, California) were mounted 6.5 cm apart on the base wall below two additional nose-pokes. Liquid reward consisting of 10 μ l evaporated milk was delivered directly to these wall-mounted nose-pokes via a single-syringe pump (New Era Pump Systems, Farmingdale, New York). Access to these response nose-pokes was restricted by a rotating, servo-controlled (Tower Hobbies, Champaign, Illinois) disc (radius, 7 cm). Rewards could be accessed from these nose-pokes only when two holes in the rotating disc were aligned with the underlying nose-pokes. Trial logic was controlled by custom software running on an Arduino Leonardo microcontroller (Ivrea, Italy).

Training. Mice were food restricted to 85–90% of their *ad libitum* body weight before training. Training consisted of multiple levels. First, mice were habituated to the test box and allowed to collect reward freely. Reward availability was signalled by the rotation of the aforementioned wall-mounted disc. The location of reward (left or right poke) was indicated by either a visual or an auditory stimulus.

For ‘attend to vision’ (visual) trials, the rewarded response poke was indicated by illumination of the LED mounted underneath it. In ‘attend to audition’ (auditory) trials, an upswing (10–14 kHz, 500 ms) indicated a reward on the left and a downswing (16–12 kHz, 500 ms) indicated a reward on the right. To facilitate discrimination learning, sweeps were initially presented in a directional manner.

Trials were given in single-modality blocks of six, with alternating block type (that is, six visual trials followed by six auditory trials; Extended Data Fig. 1, top row). The stimulus was presented until the animal collected the reward. An individual trial was terminated 20 s after reward collection and a new trial became available 5 s later. Second, mice learnt to poke to receive a reward. All other parameters remained constant. An incorrect poke had no negative consequence. By the end of this training stage, all mice collected at least 20 rewards per 30-minute session.

Third, mice were trained to initiate individual trials, allowing for the establishment of a temporal window in which they could anticipate subsequent delivery of the stimuli. For successful initiation, mice had to break the infrared beam briefly (50 ms) in the initiation poke to trigger stimulus presentation and rotation of the wall-mounted disc. Mice were informed about trial availability and modality type by brown noise (10-kHz low-pass-filtered white noise, visual trial) or blue noise (11-kHz high-pass-filtered white noise, auditory trial) delivered binaurally. At this stage, modality types were arranged in a non-conflicting block design (Extended Data Fig. 1, top row). Correct poking resulted in reward delivery, whereas incorrect poking resulted in immediate termination of the trial by disc rotation, blocking access to reward. Rewards were available for 15 s following correct poking, followed by a 5-s intertrial interval (ITI). Incorrect poking was punished with a timeout, which consisted of a 30-s ITI. Mice could not initiate new trials during an ITI. To avoid development of side preferences, the target stimulus would appear at the same location as it did on the previous trial following an incorrect response. After one week of training on this stage, mice successfully associated the target stimuli

with the appropriate reward location (Extended Data Fig. 1, top row). At this stage, directionality of sound stimuli did not affect performance.

Fourth, mice had to resolve sensory conflict. Auditory and visual target stimuli were always presented in a conflicting manner (Extended Data Fig. 1, middle row). The brown and blue noise cues indicated the modality to be selected. During a session, four different trial types were presented in blocks in repeating order: (1) three auditory trials, (2) three visual trials, (3) six conflict trials with an auditory target and (4) six conflict trials with a visual target. To prevent modality preferences, an incorrect response resulted in the repetition of the same trial type, thereby specifically increasing the block length of the trial types with weak performance. This training stage was introduced to teach mice to attend only to the target modality during a conflict trial. Over the course of this training stage (1 week), the duration of the target stimuli was successively shortened to 3 s, 1 s and 0.5 s. In parallel, the time that mice had to break the infrared barrier in the initiation nose-poke was continually increased and randomized to a final range of 0.5–0.7 s, rendering the precise presentation time of target stimuli unpredictable. Once mice performed successfully on conflict trials (Extended Data Fig. 1, middle row) the single-modality trials were removed and block length was reduced to three trials. This change in the training paradigm was made to facilitate learning of the trial-type cueing (brown and blue noise).

On the fifth and final stage of training, all block structure was removed and trial type was randomized (Extended Data Fig. 1, bottom row). We used three measures to ensure that mice followed the trial-type cueing and did not employ simple alternating strategies. In addition to computing overall accuracy (Extended Data Fig. 1, graphs on left), we quantified the number of consecutive correct trials (Extended Data Fig. 1, middle column) and calculated the fraction of correct modality switches (Extended Data Fig. 1, graphs on right). At this final stage, rewards were available for only 5 s.

Psychophysics. For experiments determining the visual detection psychometric function, the ratio between visual and auditory trials was adjusted from the typical 1/1 to 4/1 to facilitate the acquisition of a larger number of visual trials while maintaining the divided-attention nature of the task. In addition, visual stimulus duration was shortened to 0.1 s and the light was randomly displayed at one of five different intensities (0.15, 0.3, 0.6, 0.9, 1.2 lm). To establish the comparison between psychometric functions of visual-only and divided-attention trials (Fig. 1c), we trained mice that reached criterion (>70% accuracy) on the cross-modal task to perform a visual-only task. For one week, mice were trained on a visual-only task every other day; trials containing only visual target stimuli were cued by broadband white noise. Subsequently, visual-only trials were introduced into the cross-modal task at a 1/4 ratio and in a random interleaved manner. Mice were found to differentially anticipate visual-only and visual target with auditory conflict trials (Fig. 1c), whereas they continued to perform equally well on conflict trials with an auditory target (Extended Data Fig. 3).

To separate the effect of anticipating a conflicting stimulus (top-down) from the presence or absence of a distracting stimulus itself (bottom-up), we performed two experiments. In the first experiment, mice performed the cross-modal task with 70% conflict trials and 30% in which conflict was expected but auditory distraction was removed (Fig. 1d). In the second experiment, mice that had been trained on the cross-modal task had the biasing cues replaced with broadband white noise and the modality rewarded was changed on a session-by-session basis such that mice would deduce it based only on reward history (Fig. 1e).

Behavioural analysis and determination of visual detection threshold. Performance in behavioural tests was assessed based on the fraction of correct responses relative to chance level or guess rate (50%, γ). The visual detection threshold (α) and maximum performance (λ) were estimated by fitting performance across stimulation intensities with a logistic function^{29,30}:

$$F(x; \alpha, \beta, \lambda, \gamma) = \gamma + \frac{(1 - \gamma - \lambda)}{1 + \exp(-\beta(x - \alpha))}$$

where x corresponds to the five stimulus levels expressed as a percentage of maximum stimulus intensity. The fraction of correct trials as a percentage of all trials was summed across sessions and the overall performance as a function of stimulus intensity was fit using maximum likelihood estimation³⁰ implemented in the Palamedes psychophysical toolbox (<http://www.palamedestoolbox.org/>). Estimation of the distribution of the α parameter was made via non-parametric bootstrap analysis of curve fits (Fig. 1). To adjust for variable lapse rates (Extended Data Fig. 2), the fraction of correct trials was normalized so that the minimum and maximum performance rates corresponded to 50% and 100%, respectively³¹. Curve fitting and estimation of the α parameter then proceeded as described above. Model selection for the number of psychophysical parameters was based on the Akaike information criterion^{20,32}.

Optogenetics in behaviour. For experiments with optical stimulation (Fig. 2), testing conditions were equivalent to the final stage of training. Laser trains of

either blue (for Chr2 activation) or yellow (for eNpHR3.0 activation) light consisting of 50-Hz 18-ms pulses (90% duty cycle) at an intensity of 5–6 mW (measured at the tip of the optic fibres) were delivered on every other trial. On laser trials, stimulation occurred either during the anticipatory period (0.5–0.7 s) or during stimulus presentation (0.5 s). Because behaviour and recording systems were automated and stimulus sequence and optogenetic manipulations varied on a trial-by-trial basis, researchers were not blinded to the conditions. In the case of multiple sequential pharmacological or optogenetic manipulations in the same animals, tests were performed in a predefined, pseudorandom order. For comparisons of multiple groups, Kruskal–Wallis one-way analysis of variance was used to assess variance across groups before pairwise comparisons. Power analysis based on effect size estimates was used to determine sample size required for statistical significance with a power of $\beta = 0.7$; more than three samples were required to detect significant differences.

Electrophysiology and optical chloride measurements in behaviour. For combined TRN recordings with optogenetic PFC disruption (Fig. 3d, e), laser trains of blue light (as described earlier) were delivered during the anticipatory period on every other trial. For electrophysiological recordings of LGN units and fibre photometry measurements, visual stimuli were presented through illumination of diffusion-coated wide-angle 3-mm flat-top LED lights (LightHouseLEDs, Washington), fixed directly on the head of the mouse and centred 8 mm from the eyes. The LEDs mounted on the base wall of the behavioural box were turned off in this condition. These changes allowed emitted light to activate $\sim 150^\circ$ of the visual angle when the eye is centred at rest³³.

Viruses. For retrograde optogenetic tagging and TRN manipulation, FuGB2-pseudotyped retrograde lentiviruses (RG-LV) were used as described previously²⁰. visTRN neurons were labelled through injection (0.4–0.6 μ l) of RG-LV-EF1 α -DIO-ChR2-GFP (for activation) or RG-LV-EF1 α -DIO-eNpHR3.0-eYFP (for inactivation) into the primary visual thalamus (anterior–posterior (A–P), -2.1 mm; medial–lateral (M–L), ± 2 mm; dorsal–ventral (D–V), 2.5 mm) using a quintessential stereotactic injector (QSI, Stoelting, Wood Dale, Illinois). Coordinates are referenced to Bregma. For combined electrophysiological recordings with optogenetic PFC disruption during behaviour, 0.4μ l AAV2-hSyn-DIO-ChR2-GFP (titre, 10^{12} vector core per ml) was injected into the PFC (A–P, 2.6 mm; M–L, ± 0.25 mm; D–V, -1.25 mm). To measure chloride flux in the LGN, the transgene *SuperClomeleon* (gift from G. J. Augustine²⁵) was cloned into the AgeI and EcoRI restriction sites of an AAV-hSyn-SSFO-eYFP plasmid to obtain AAV-hSyn-SuperClomeleon. The SuperClomeleon recombinant AAV was packaged as serotype 2 (University of North Carolina, Vector core facility; titre, 10^{12} vector core per ml) and 0.6 – 0.7μ l virus was injected into the visual thalamus. Following injections, mice were allowed to recover for 2–4 weeks to allow for virus expression.

Optic fibre implant experiments. Mice were anaesthetized using 1% isoflurane and mounted on a stereotactic frame. For cortical inactivation experiments, up to three pairs of 4–5-mm-long optic fibres (Doric Lenses, Quebec, Canada) were inserted bilaterally to target up to three different brain areas per mouse (prelimbic cortex, 2.6 mm A–P, ± 0.25 mm M–L, -1.25 mm D–V; primary visual cortex, -3.5 mm A–P, ± 2.50 mm M–L, -0.50 mm D–V; primary auditory cortex, -2.8 mm A–P, ± 4.00 mm M–L, -2.00 mm D–V; AAC, 0.5 mm A–P, ± 0.25 mm M–L, -1.00 mm D–V; lateral OFC, 2.6 mm A–P, ± 1.50 mm M–L, -2.00 mm D–V; primary visual thalamus, -2.1 mm A–P, ± 2.00 mm M–L, -2.50 mm D–V; visTRN, -1.6 mm A–P, ± 2.20 mm M–L, -3.00 mm D–V). Two or three stainless steel screws were implanted into the skull to anchor the implant and were fixed with dental cement. Animals were allowed to recover and training resumed one week later. For Chr2 activation a 473-nm laser was used and a 561-nm laser was used for eNpHR3.0 activation (Omicron-Laserage, Dudenhofen, Germany).

Drive construction. Custom drive housings were designed using 3D CAD software (SolidWorks, Concord, Massachusetts) and printed in Accura 55 plastic (American Precision Prototyping, Tulsa, Oklahoma) as described previously^{20,34}. Prior to implantation, each drive was loaded with 8–12 independently movable microdrives carrying up to 3 nichrome (12.5μ m) and/or tungsten (25μ m) stereotrodes (California Fine Wire Company, Grove Beach, California). Stereotrodes were pinned to custom-designed 32- or 64-channel electrode interface boards (EIB; Sunstone Circuits, Mulino, Oregon) along with a common reference wire (A–M systems, Carlsborg, Washington). For optogenetic tagging, an optical fibre was embedded adjacent to the stereotrode array. In these cases, the optic fibre extended 3.5 mm from the base of the drive so that it could be stereotactically positioned above the TRN during implantation. Targeting of the TRN or LGN was achieved by guiding stereotrodes and optic fibres through a square array of polyimide sleeves attached to the base of the drive body.

Drive and fibre implantation for optical activity measurements. Prior to surgical implantation, mice were anaesthetized with 1% isoflurane and placed in a stereotactic frame. Stainless steel screws were implanted into the skull to provide

electrical and mechanical stability for the drives. For drive implantations, craniotomies (~ 3 mm \times 2 mm) were drilled, centred at -2 mm A–P and 2.5 mm M–L for TRN recordings (15° angled implantation) and at -2.3 mm A–P and -2.5 mm M–L for LGN recordings. The dura mater was carefully removed and drives were centred at the craniotomy coordinates using a custom stereotactic arm. Drive bodies were slowly lowered into the craniotomy until stereotrode tips were $\sim 500 \mu$ m below the cortical surface and optical fibres were positioned just above the TRN (2.5 mm D–V). For fibre-photometry-based optical recording, low-internal-fluorescence optic fibres (400 μ m diameter) (Doric Lenses, Canada) were implanted just dorsal to the LGN (-2.2 mm A–P, 2.15 mm M–L and 2.6 mm D–V) following virus injection.

Electrophysiological recordings. After mice had recovered from implantation surgery, recordings were made using a Neuralynx multiplexing digital recording system (Neuralynx, Bozeman, Montana). Signals were acquired using a 32- or 64-channel digital headstage connected to the implanted EIB. Signals from each electrode were amplified, filtered between 0.1 Hz and 9 kHz and digitized at 30 kHz. Local field potential signals were obtained from a single wire per stereotrode. Following implantation, stereotrode sets were incrementally lowered from the cortex into the target thalamic structure over the course of 1–2 weeks (Extended Data Fig. 5). Spike sorting was performed offline following acquisition based on relative spike amplitude and energy within electrode pairs using the MClust toolbox (<http://redishlab.neuroscience.umn.edu/mclust/MClust.html>). Following manual clustering, cross-correlation and autocorrelation analyses were used to confirm adequate separation.

Optogenetically tagged visTRN units were identified based on Chr2-mediated response to stimulation using a 473-nm analogue-modulated laser (Omicron-Laserage, Dudenhofen, Germany)²⁰. Laser light was delivered by a $200\text{-}\mu$ m optic fibre targeted to the TRN (Extended Data Fig. 5) connected to a fibre optic patch cord ($200\text{-}\mu$ m core, Doric Lenses, Quebec, Canada). The laser intensity was set at ~ 8 mW optical output power measured at the patch cord terminus. Fibres were polished before implantation so that the power at the tip was $\geq 50\%$ maximum, resulting in ~ 4 – 5 mW laser light being delivered to the brain. Only neurons that showed clear transient responses to laser stimulation were included in the analysis.

Analysis of firing rate. Changes in firing rate during task performance were assessed for 138 identified visTRN neurons recorded from four animals and 119 LGN neurons in two animals. Peri-event time histograms aligned to trial initiation and to stimulus presentation were computed using a 5-ms bin width for individual neurons in each recording session⁴. Separate histograms were created for correct and incorrect trials within auditory and visual target stimuli and convolved with a Gaussian kernel (8 ms half-width at half-height) to create a spike density function^{35,36}. The average firing rate across trials was determined during the anticipation window before stimulus presentation. The evoked response amplitude was estimated by averaging the firing rate within a 100-ms window starting 20 ms after stimulus onset. Window duration was chosen based on the latency-to-peak response for point stimuli in the mouse LGN³⁷.

For normalized rate changes in TRN neurons, firing rates during the attentional window in each trial were compared with the baseline firing rate (5-s window, 0.5 s before task initiation). Statistical comparison of firing rate changes was used to identify neurons with significant task-associated changes in firing rate via non-parametric comparison of firing rate during the attentional window and the baseline period³⁸. The test statistic (W) was calculated based on ranking of all trials (N) and comparison was performed using the sign function (sgn):

$$W = \left| \sum_j^N [\text{sgn}(x_{2,j} - x_{1,j}) \cdot R_j] \right|$$

where x_1 and x_2 were the attentional window and baseline firing rate, respectively, and R denotes the rank. The threshold for significance was set at 0.05 and significantly modulated units were defined as neurons in which the test statistic was less than the critical value for the sample size ($W_{\alpha(0.05)}$). Comparison of firing rates across trial types (for example, visual versus auditory correct) was performed using the Wilcoxon rank-sum test. Homogeneity of variance for firing rates across conditions was determined using the Fligner–Killeen test of homoscedasticity.

Analysis of visual evoked potentials. Visual evoked potentials (VEPs) were computed from the broadband LGN local field potential (LFP; 0.1 Hz– 10 kHz). The particular stereotrode used for VEP analysis in behaviour was selected based on the amplitude of responses in post-task recordings during which there were many more trials included. Task-related VEPs were averaged during correct auditory and visual trials across recording sessions. To determine peak response, the lowest negative-potential offsets associated with the visual response³⁹ (0 – 250 -ms window) were identified on a trial-by-trial basis. Signals from individual trials were smoothed with a 25-ms half-width filter over the response window before obtaining the peak offset⁴⁰.

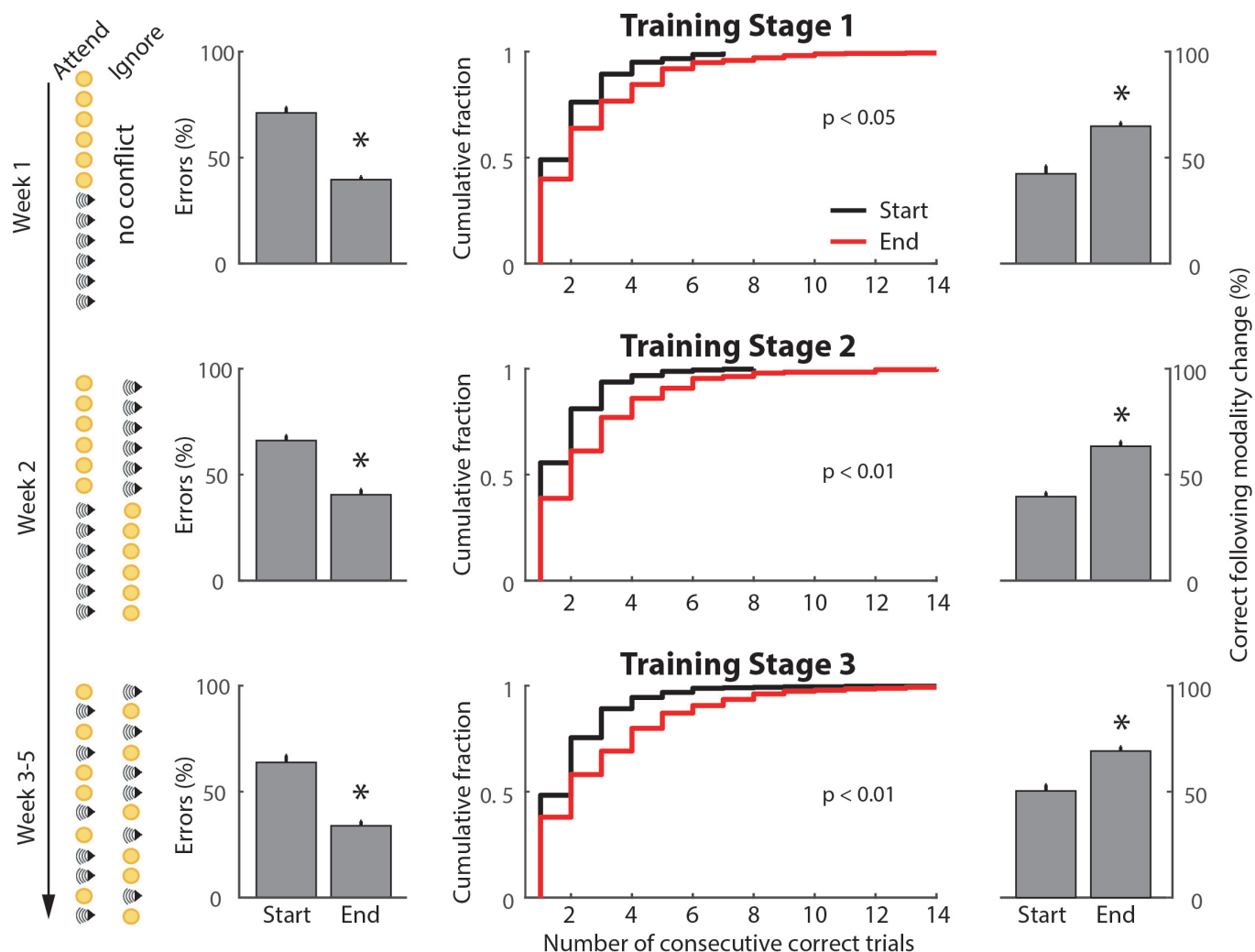
Histology. Mice were euthanized and transcardially perfused with PBS followed by 4% paraformaldehyde. Brains were dissected, post-fixed overnight at 4 °C and sectioned using a vibratome (LEICA, Buffalo Grove, Illinois).

For GFP enhancement, immunofluorescent staining was carried out on 50- μ m-thick sections using chicken anti-GFP (1/1,000, GFP-1020, Aves). Sections were incubated overnight with primary antibody in PBS-T (10% normal goat serum and 0.05% Tween20) at 4 °C. Detection of primary antibodies was carried out with Alexa-Fluor-conjugated secondary antibodies (1/1,000, A-11039, Invitrogen). All sections were imaged on a Zeiss LSM510 META confocal microscope.

Fibre-photometry-based optical chloride measurements. FRET-based measurement of chloride was performed during behaviour using a custom-designed fibre photometry system²⁴. A fibre-coupled LED (Thorlabs, Newton, New Jersey) light source, filtered using a 434-nm clean-up filter (MF434-17 Thorlabs, Newton, New Jersey), was used for CFP excitation. Excitation light was split via a long-pass dichroic mirror (DMLP425, Thorlabs, Newton, New Jersey) and coupled to a 400- μ m, 0.48-NA (pharmacology) optic patch cord (Doric lenses, Canada) linked to a 400- μ m chronically implanted optical fibre. Excitation and emission light were conveyed by a single patch cord linking the fibre photometry system to the implanted fibre. SuperClomeleon CFP and YFP emissions²⁵ were separated using a single-edge beam splitter (FF511-Di01, Semrock, Rochester, New York). Each emission wavelength was independently focused onto a separate femtowatt silicon photoreceiver (Newport, Irvine, California) using custom optics (12.7-mm focal length plano-convex lens mounted in Thorlabs SM1NR05 lens tube). The light signal was digitized and recorded using a TDT signal acquisition system (Tucker-Davis Technologies, Alachua, Florida). Signal bandwidth was limited to <750 Hz based on the photoreceiver response characteristics. The fluorescence ratio was calculated across the recording period. To minimize the effect of slow fluctuations, normalized delta fluorescence (df/F) was calculated for evoked responses relative to the baseline fluorescence level before each event (1-s window). Traces were smoothed with a convolution filter (50 ms half-width). Peak response (Extended Data Fig. 10) was estimated as the minimum over a 500-ms window following stimulus onset. For pharmacological activation of GABA_A receptors with 4,5,6,7-tetrahydroisoxazolo(5,4-c)pyridin-3-ol (THIP), baseline fluorescence was estimated over 5 minutes before injection. For visual stimulation, light pulses of 100-ms duration were displayed to the ipsi- or contralateral side of the recorded LGN.

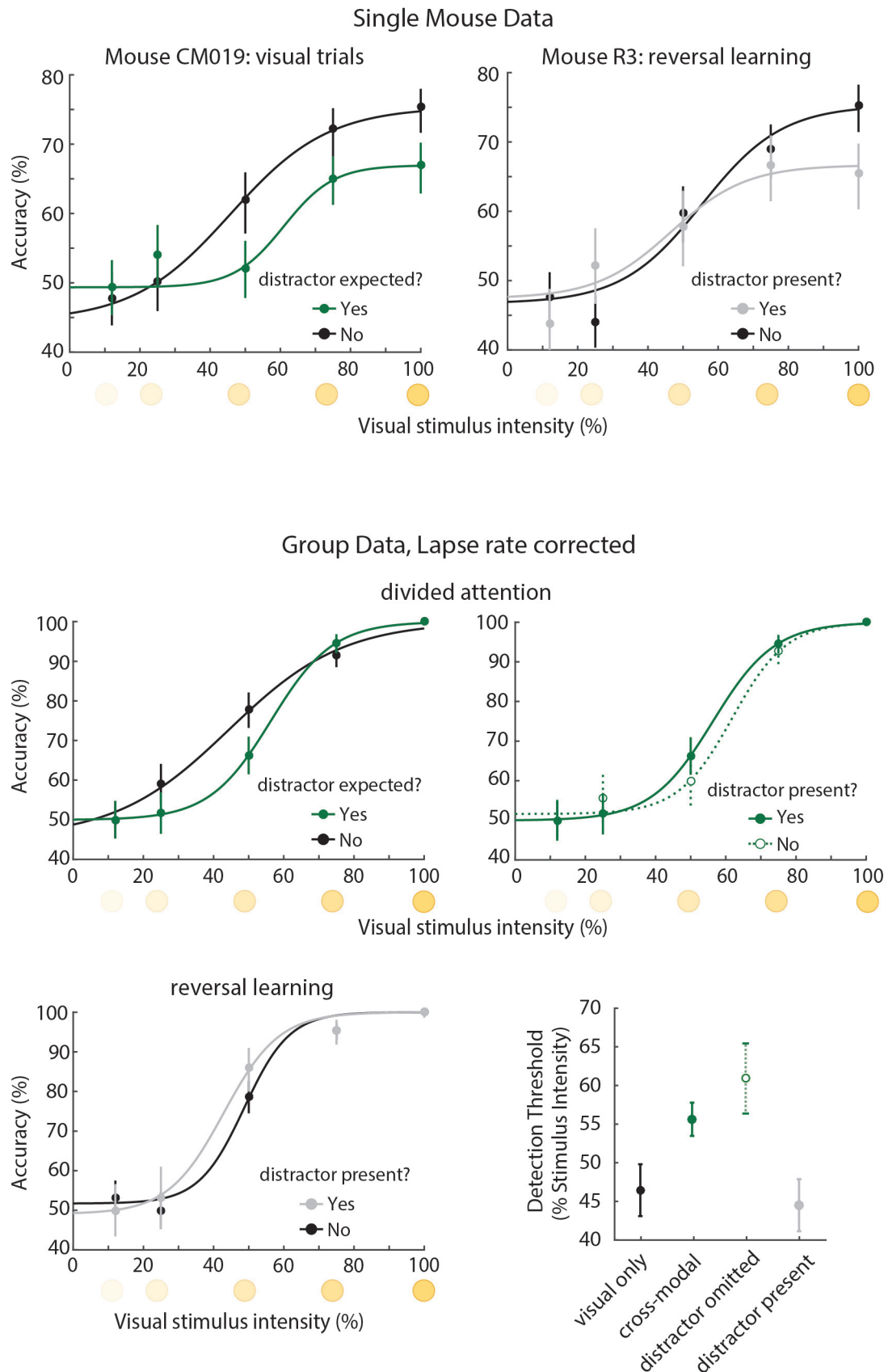
Effects of the GABA_A receptor antagonist flumazenil on visual evoked responses were quantified by comparing the average peak response from 5 minutes before injection (baseline) to one within a 5-minute time window around the maximal response suppression (maximal drug effect) and at the end of the recording session (recovery, at least 100 min after injection). For optogenetic manipulations of frontal cortical structures, smaller-diameter patch cords (200 μ m, 0.37 NA) were used to allow movement and prevent tangling. For these recordings, power analysis was performed to determine sample size required to detect significant differences with a power of $\beta = 0.7$ based on the observed differential signal in correct auditory and visual trials under baseline conditions. Analysis indicated that more than four independent samples would be required to detect a change in these differential responses.

29. Ress, D. & Heeger, D. J. Neuronal correlates of perception in early visual cortex. *Nature Neurosci.* **6**, 414–420 (2003).
30. Levitan, C. A., Ban, Y. H., Stiles, N. R. & Shimojo, S. Rate perception adapts across the senses: evidence for a unified timing mechanism. *Sci. Rep.* **5**, 8857 (2015).
31. Mareschal, I., Calder, A. J., Dadds, M. R. & Clifford, C. W. Gaze categorization under uncertainty: psychophysics and modeling. *J. Vis.* **13**, 18 (2013).
32. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction* 2nd edn, Ch. 7 (Springer, 2009).
33. Wallace, D. J. *et al.* Rats maintain an overhead binocular field at the expense of constant fusion. *Nature* **498**, 65–69 (2013).
34. Brunetti, P. M. *et al.* Design and fabrication of ultralight weight, adjustable multi-electrode probes for electrophysiological recordings in mice. *J. Vis. Exp.* **91**, e51675 (2014).
35. Fries, P., Neuenschwander, S., Engel, A. K., Goebel, R. & Singer, W. Rapid feature selective neuronal synchronization through correlated latency shifting. *Nature Neurosci.* **4**, 194–200 (2001).
36. Szucs, A. Applications of the spike density function in analysis of neuronal firing patterns. *J. Neurosci. Methods* **81**, 159–167 (1998).
37. Piscopo, D. M., El-Danaf, R. N., Huberman, A. D. & Niell, C. M. Diverse visual features encoded in mouse lateral geniculate nucleus. *J. Neurosci.* **33**, 4642–4656 (2013).
38. De Araujo, I. E. *et al.* Neural ensemble coding of satiety states. *Neuron* **51**, 483–494 (2006).
39. Ridder III, W. H. & Nusinowitz, S. The visual evoked potential in the mouse—origins and response characteristics. *Vision Res.* **46**, 902–913 (2006).
40. Izaki, Y., Fujiwara, S. E. & Akema, T. Rat prefrontal response and prestimulation local field potential power in vivo. *Neuroreport* **19**, 255–258 (2008).



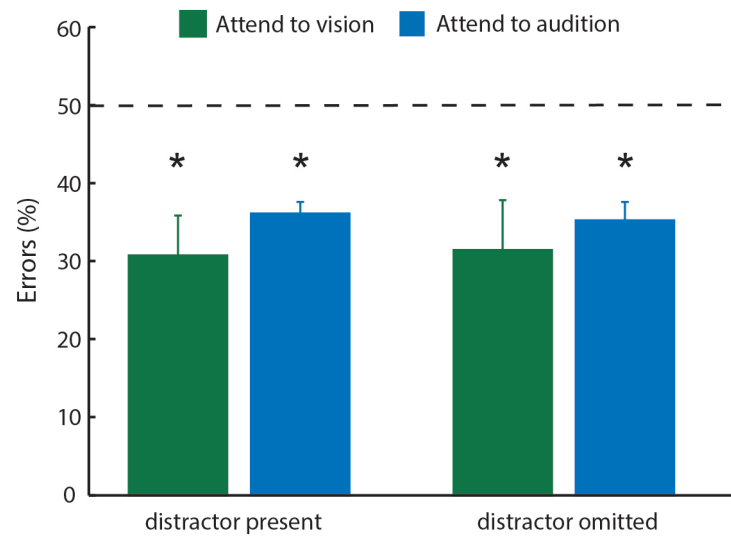
Extended Data Figure 1 | Cross-modal task training and performance validation. Quantification of performance across training stages for the cross-modal task. The trial sequence for each training stage is indicated on the left. Improved performance was observed in the last three days of training relative to the first three for each stage. Bar graphs on the left (column 1) show the

reduction in the error fraction ($n = 15$ mice, $*P < 0.05$, Wilcoxon rank-sum test), column 2 shows the number of consecutive correct responses (P -values shown, Kolmogorov-Smirnov test) and bar graphs on the right (column 3) show the probability of a correct response following a modality shift ($*P < 0.05$, Wilcoxon rank-sum test).



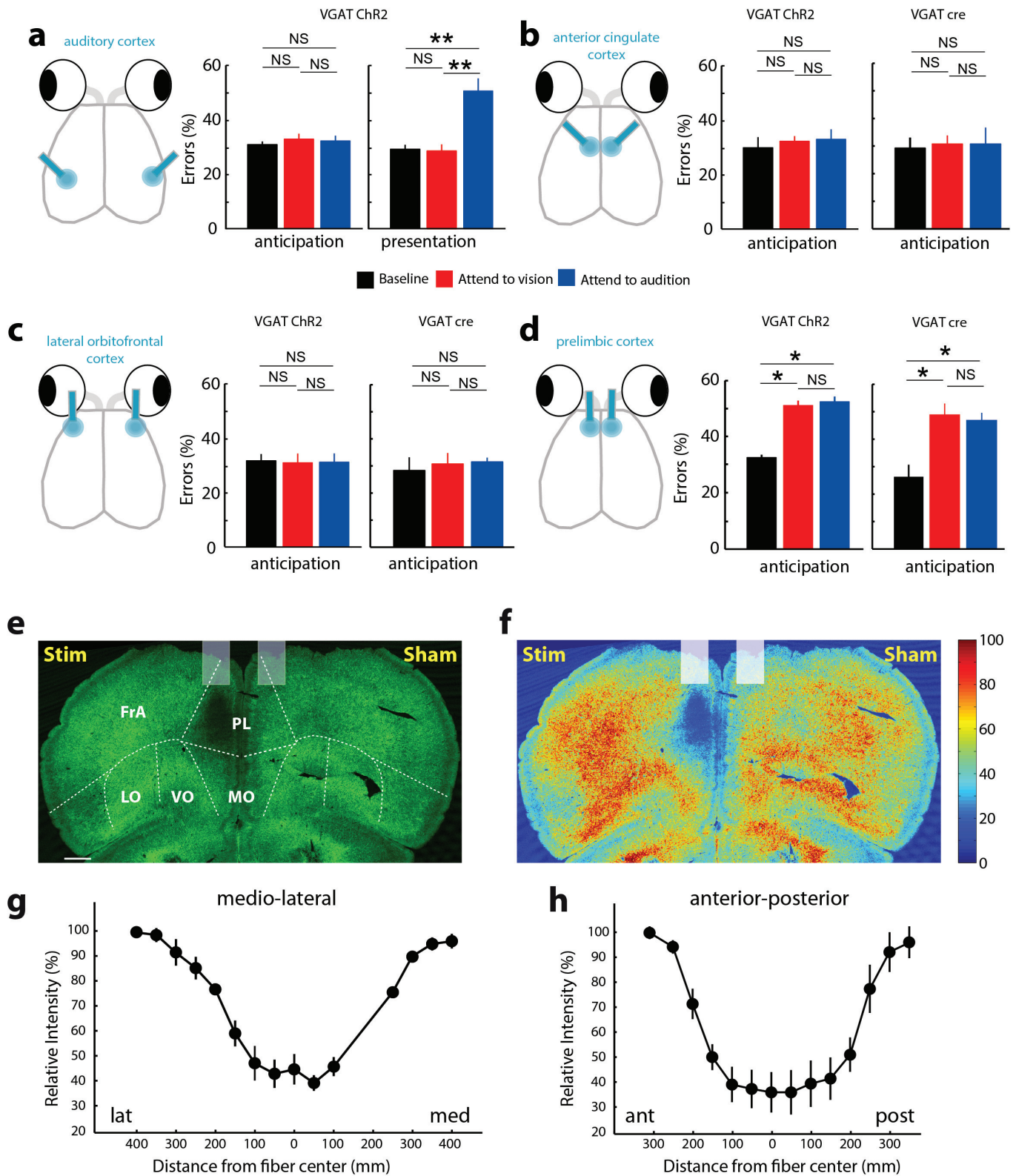
Extended Data Figure 2 | Effects of cross-modal divided attention in the mouse. Top row, single-mouse examples of visual detection performance during cross-modal divided attention and reversal learning. Comparison of performance under visual-only (black) and cross-modal (green) conditions are shown on the left. Although neither condition contained sensory conflict, the mere expectation of one increased detection threshold (≥ 124 trials per

condition). Detection threshold was not affected by the presence of an auditory distractor during reversal learning (≥ 90 trials per condition), as shown on the right. Middle and bottom rows, group data normalized to peak performance (lapse rate), showing that the effects of divided attention on detection threshold were persistent. Bootstrap estimation of visual detection threshold shows a similar pattern as data in Fig. 1 (error bars are 95% confidence intervals).



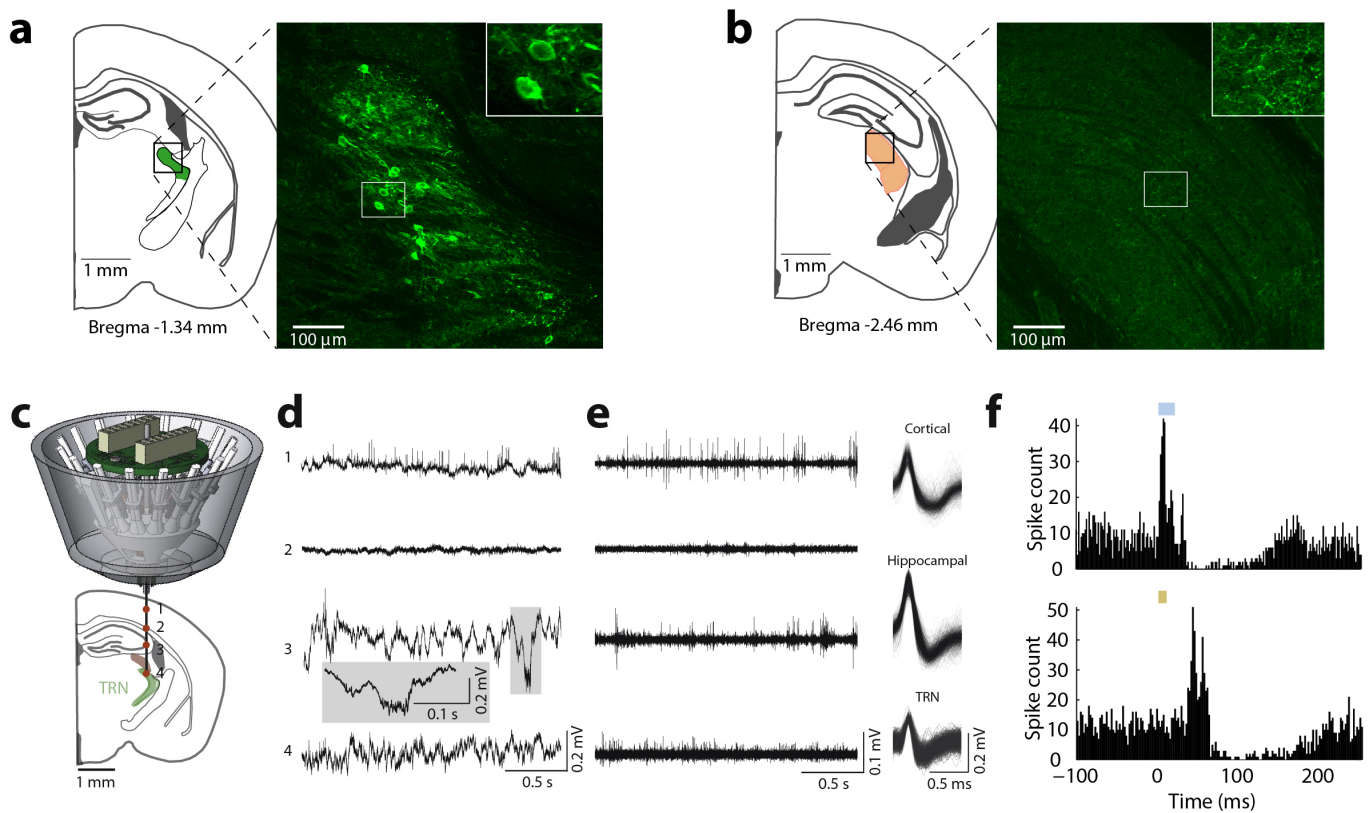
Extended Data Figure 3 | Comparable performance on trial types and intact overall auditory performance despite auditory stimuli being eliminated on a subset of ‘attend to vision’ trials. Left, performance was comparable on

auditory and maximum-intensity visual trials ($n = 4$ mice, same as in Fig. 1d). Right, mice exhibited comparable overall performance when auditory stimuli were eliminated from a subset of ‘attend to vision’ trials.



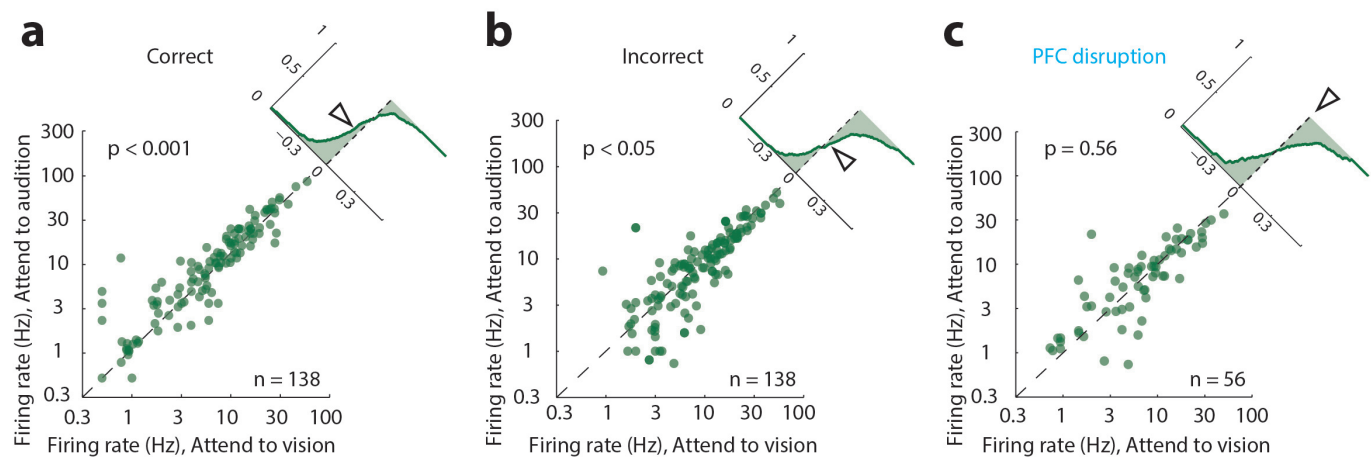
Extended Data Figure 4 | Region- and timing-specific effects of optogenetic manipulation on cross-modal task performance. **a**, Optogenetic disruption of auditory cortex during target stimulus anticipation disrupted performance specifically for auditory trials ($n = 4$ mice, $**P < 0.01$, Wilcoxon rank-sum test). Disruption of AAC (**b**) or lateral OFC (**c**) in VGAT-ChR2 mice or following localized injection of a Chr2-expressing virus did not affect performance ($n = 4$ mice (2 VGAT-ChR2 and 2 VGAT-Cre), 4 sessions per manipulation). **d**, In contrast, inactivation of prelimbic (PL) cortex led to robust reduction in performance in both types of manipulation ($n = 8$ mice

(4 VGAT-ChR2 and 4 VGAT-Cre), $*P < 0.05$ Wilcoxon rank-sum test). **e–h**, Photobleaching experiment to quantify the spread of laser light. A coronal section (**e**) shows GFP bleaching following two-hour exposure to laser stimulation (6 mW, 50 Hz, 90% duty cycle). **f–h**, Fluorescence intensity quantification shows that the extent of light spread is limited to 300 μ m around the tip of the optic fibre ($n = 3$ mice). Ant, anterior; FrA, frontal association cortex; lat, lateral; LO, lateral orbitofrontal cortex; med, medial; MO, medial orbitofrontal cortex; post, posterior; Stim, stimulation; Sham, sham surgery control; VO, ventral orbitofrontal cortex. Scale bar in **e**, 200 μ m.



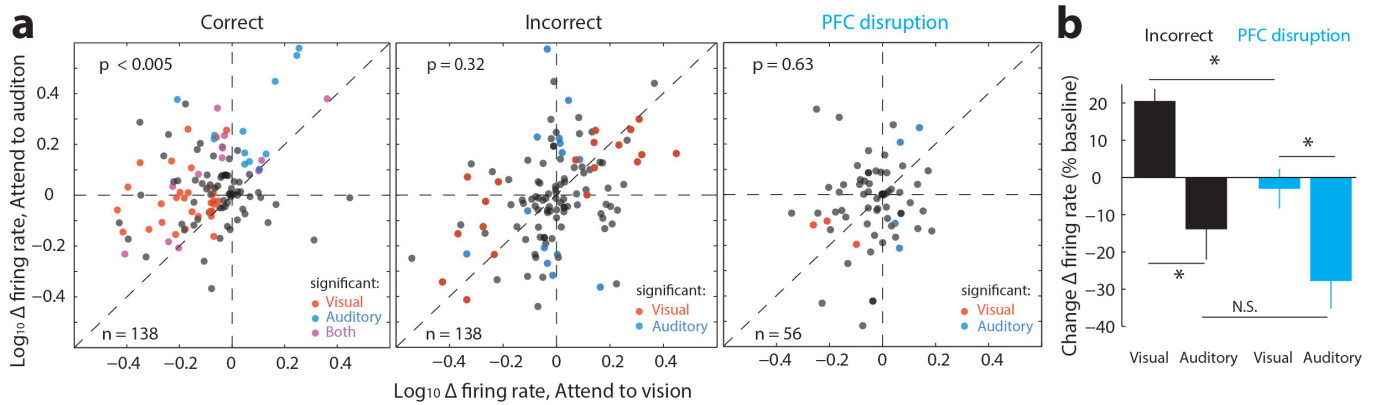
Extended Data Figure 5 | Independently adjustable, multi-electrode recording of visTRN neurons. **a, b**, Injection of DIO-ChR2-eYFP retrograde lentivirus into LGN labels visTRN neurons but not LGN interneurons. **a**, The histological image is the maximal projection of four 2- μ m confocal planes showing labelling of visTRN neurons (inset shows a zoom view of cell bodies) approximately 1.34 mm posterior to Bregma. **b**, Image as in **a**, but from LGN of the same animal, approximately 2.46 mm posterior to Bregma (inset shows a zoom view of terminals). **c**, Schematic of independently adjustable multi-electrode drive. **d**, An example of activity recorded from different depths during adjustment. Distinct patterns of physiological activity are observed along the

trajectory in the broadband local field potential signal (0.1 Hz–32 kHz). The numbers correspond to different recording sites (marked by the red dots on the schematic in **c**). **e**, High-pass-filtered signals (600 Hz–10 kHz) showing spiking activity, with isolated clustered units showing distinguishable waveform characteristics in distinct structures. **f**, Example peri-event time histograms of ChR2-mediated visTRN response. Top, response to laser activation (473 nm, ~4 mW, stimulation, 20 ms). Bottom, response to visual stimuli (10-ms pulse). Blue and orange blocks indicate laser and visual stimulation, respectively.



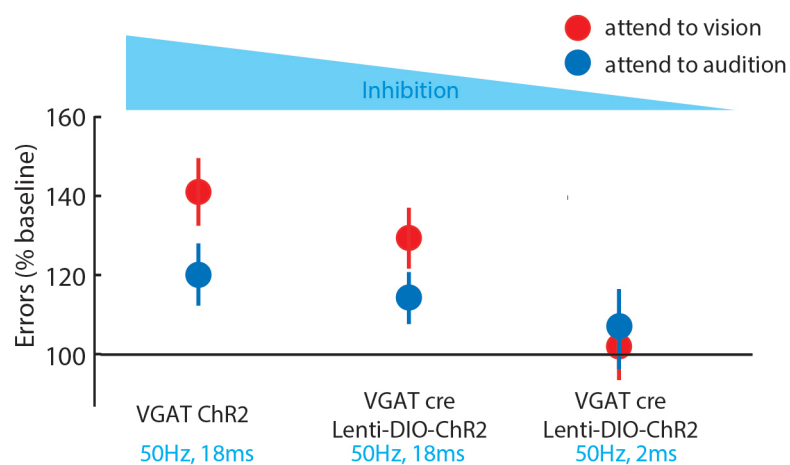
Extended Data Figure 6 | Distinct changes in visTRN firing rate during natural errors compared to errors due to PFC disruption. **a–c**, Scatter plots showing the change in absolute firing rate for visTRN neurons for correct (**a**), incorrect (**b**) or disrupted-PFC trials (**c**). Insets show the cumulative probability plot of separation from the unity line (no change). Although correct trials had a lower firing rate in ‘attend to vision’ than in ‘attend to audition’ trials ($n = 138$,

$P < 0.001$ Wilcoxon signed-rank test), this pattern was reversed for incorrect trials ($n = 138$, $P < 0.05$, Wilcoxon signed-rank test); this suggests that perhaps the animal was attending to the wrong modality. This reversal was not observed in trials with PFC disruption (despite mouse performance being at chance level).



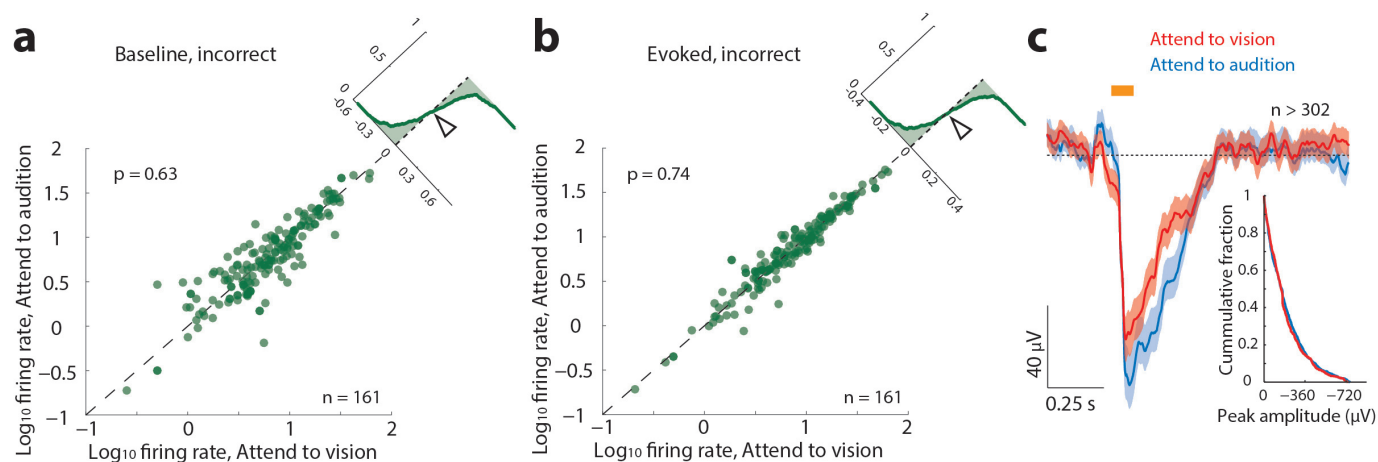
Extended Data Figure 7 | The effect of PFC disruption on visTRN activity is distinct from naturally occurring errors. **a**, Scatter plots of response from visTRN neurons, comparing the modulation of their firing rate (change from baseline) under the two distinct anticipatory conditions. Each sample is a single cell. Colours denote significance reached for each cell on a trial-by-trial basis (red, visual; blue, auditory; purple, both; rank-sum-test comparison to baseline). Note that in correct performance ($n = 138$, 4 mice, $P < 0.005$, Wilcoxon signed-rank test), 'attend to vision' resulted in a negative shift and

'attend to audition' resulted in a positive shift, consistent with examples shown in Fig. 3. During naturally occurring error trials, the modulation is partially reversed for both trial types, suggesting that at least a subset of errors are the result of attending to the wrong modality. In contrast, PFC disruption ($n = 56$ cells, 2 mice) resulted in a weaker, non-uniform effect ('attend to visual' trials are less affected). **b**, Quantification of effects seen in **a**. N.S., not significant.



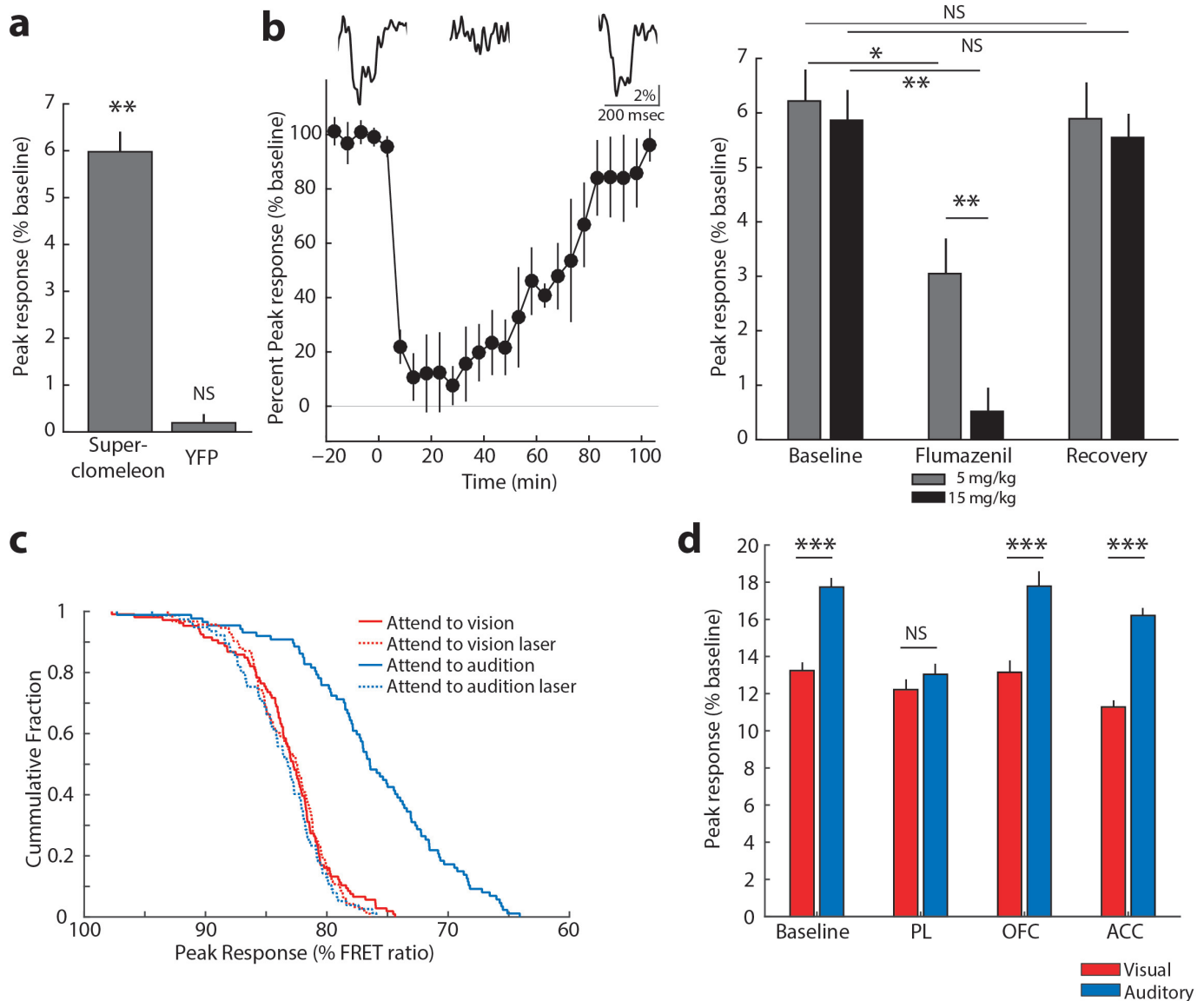
Extended Data Figure 8 | The magnitude of behavioural disruption co-varies with the strength of optogenetic manipulation of the LGN or visTRN. Activation of inhibitory terminals in the LGN with a 90% duty cycle laser (Fig. 2) resulted in maximal disruption of cross-modal performance. Activating

visually labelled TRN with identical stimulation parameters resulted in a quantitatively lower behavioural effect. Reducing the duty cycle of visTRN stimulation to 10% resulted in no effect on accuracy, as shown previously⁴.



Extended Data Figure 9 | Attentional modulation by LGN is not observed on error trials. **a, b**, No significant difference was observed in the average firing rate of LGN neurons during stimulus anticipation ($P = 0.63$, Wilcoxon signed-rank test, $n = 161$ cells, 4 mice) or presentation ($P = 0.74$, Wilcoxon

signed-rank test, $n = 161$) among trial types when behavioural outcomes were incorrect. **c**, Similar effects were observed for VEPs (visual, $n = 324$ trials; auditory, $n = 302$ trials; 4 mice).



Extended Data Figure 10 | Light-evoked fast transients from chloride photometry measured in the LGN are GABA_A-receptor dependent and sensitive to visTRN and prelimbic inactivation in the cross-modal task. **a**, Peak SuperClomeleon FRET- and YFP-control responses to light stimuli (50 ms, 0.1 Hz) delivered to the eye contralateral to the recorded LGN ($n > 90$ trials from 3 mice for SuperClomeleon and from 4 mice for YFP, $***P < 0.001$, Friedman test). **b**, Chloride photometry transients are sensitive to the GABA_A receptor antagonist flumazenil in a dose-dependent manner. Left, intraperitoneal injection of 15 mg kg^{-1} flumazenil resulted in a 90% peak reduction of light-evoked chloride photometry responses, which recovered over the course of 90–100 min as predicted by flumazenil pharmacokinetics. Insets show example traces of single events recorded during baseline, peak suppression and recovery. Right, quantification of the maximal suppressive effects and recovery of 5 mg kg^{-1} and 15 mg kg^{-1} flumazenil on chloride

photometry responses ($n > 90$ trials from 3 mice, $*P < 0.05$, $**P < 0.01$, Friedman test). **c**, Cumulative distributions of unitary visual-evoked SuperClomeleon FRET peaks in response to light stimuli in the cross-modal task. Under baseline conditions, 'attend to audition' trials exhibited significantly larger amplitudes than 'attend to vision' trials, consistent with average data in Fig. 5f. Optogenetic silencing of visTRN neurons eliminated the difference between trial types and resulted in peak amplitudes comparable to baseline 'attend to vision' trials ($n = 3$ mice, $P < 0.005$ for 'attend to audition' trials vs all other trial types, Kolmogorov–Smirnov statistics with Bonferroni correction). **d**, Combined optogenetic and chloride photometry inactivation of different frontal cortical regions in the LGN while mice performed the cross-modal task. Only PL inactivation eliminates differential inhibition between visual and auditory trials ($n = 6$ mice, $***P < 0.001$, Wilcoxon rank-sum test).

CMT2D neuropathy is linked to the neomorphic binding activity of glycyl-tRNA synthetase

Weiwei He^{1*}, Ge Bai^{2*}, Huihao Zhou¹, Na Wei¹, Nicholas M. White², Janelle Lauer³, Huaqing Liu⁴, Yi Shi¹, Calin Dan Dumitru¹, Karen Lettieri², Veronica Shubayev⁴, Albena Jordanova⁵, Velina Guerguelcheva⁶, Patrick R. Griffin³, Robert W. Burgess⁷, Samuel L. Pfaff² & Xiang-Lei Yang¹

Selective neuronal loss is a hallmark of neurodegenerative diseases, which, counterintuitively, are often caused by mutations in widely expressed genes¹. Charcot-Marie-Tooth (CMT) diseases are the most common hereditary peripheral neuropathies, for which there are no effective therapies^{2,3}. A subtype of these diseases—CMT type 2D (CMT2D)—is caused by dominant mutations in *GARS*, encoding the ubiquitously expressed enzyme glycyl-transfer RNA (tRNA) synthetase (GlyRS). Despite the broad requirement of GlyRS for protein biosynthesis in all cells, mutations in this gene cause a selective degeneration of peripheral axons, leading to deficits in distal motor function⁴. How mutations in GlyRS (GlyRS^{CMT2D}) are linked to motor neuron vulnerability has remained elusive. Here we report that GlyRS^{CMT2D} acquires a neomorphic binding activity that directly antagonizes an essential signalling pathway for motor neuron survival. We find that CMT2D mutations alter the conformation of GlyRS, enabling GlyRS^{CMT2D} to bind the neuropilin 1 (Nrp1) receptor. This aberrant interaction competitively interferes with the binding of the cognate ligand vascular endothelial growth factor (VEGF) to Nrp1. Genetic reduction of Nrp1 in mice worsens CMT2D symptoms, whereas enhanced expression of VEGF improves motor function. These findings link the selective pathology of CMT2D to the neomorphic binding activity of GlyRS^{CMT2D} that antagonizes the VEGF–Nrp1 interaction, and indicate that the VEGF–Nrp1 signalling axis is an actionable target for treating CMT2D.

CMT diseases are a group of inherited disorders that specifically affect the peripheral nervous system and are characterized by progressive weakness and atrophy in the hands and feet^{2,3}. Recent progress in neurogenetic studies has uncovered aminoacyl-tRNA synthetases as the largest gene family implicated in CMT. Among them, *GARS*, encoding GlyRS, was the first member identified, mutations in which cause a dominant axonal form of CMT (CMT2D)⁴. The canonical function of this evolutionarily ancient enzyme is to catalyse the ligation of glycine to the 3' end of its cognate tRNA as the first step of protein synthesis. Interestingly, emerging evidence has revealed that GlyRS in multicellular organisms, like several other tRNA synthetase family members, has acquired the ability to be secreted from cells and as an extracellular protein can influence cell signalling^{5–9}.

More than a dozen missense *GARS* mutations (GlyRS^{CMT2D}) have been found in CMT2D patients, with varying degrees of genetic evidence for disease association^{10–13}. Among them, three mutations (E71G, L129P and G240R) are the most tightly linked to the disease¹¹ (Fig. 1a). Spontaneous and *N*-ethyl-*N*-nitrosourea (ENU)-induced missense mutations in mouse *Gars* also cause CMT2D-like phenotypes^{14,15} (Fig. 1a). These dominant mutations are found throughout the primary sequence of GlyRS, with some affecting the aminoacylation enzyme activity whereas others do not¹¹. Mice with a heterozygous

deletion of the *Gars* gene and a 50% reduction in glycyl-tRNA synthetase activity are normal¹⁵. Furthermore, overexpression of the wild-type GlyRS (GlyRS^{WT}) in a CMT2D-disease mouse fails to rescue the neuropathy¹⁶. These genetic experiments indicate that CMT2D may arise from an abnormal activity gained by GlyRS^{CMT2D} rather than a general defect in tRNA aminoacylation as initially suspected.

In common with most class II tRNA synthetases, GlyRS functions as a dimer for aminoacylation. Interestingly, despite being dispersed in three separate domains of GlyRS, all known CMT2D-causing mutations are located near the dimer interface in the GlyRS crystal structure¹⁷. We found that five different human mutations associated with CMT2D caused a conformational opening in GlyRS that exposes new protein surfaces to solution¹⁷ (Fig. 1b). To test if this conformational change also occurs in GlyRS^{CMT2D} linked to CMT-like phenotypes in mice, we performed hydrogen–deuterium exchange labelling on GlyRS(P234KY) and found that the mouse mutation likewise opens new surfaces of the GlyRS protein to solution (Fig. 1c and Extended Data Fig. 1). These findings suggest that the abnormal ‘opening’ is shared by many CMT2D mutants.

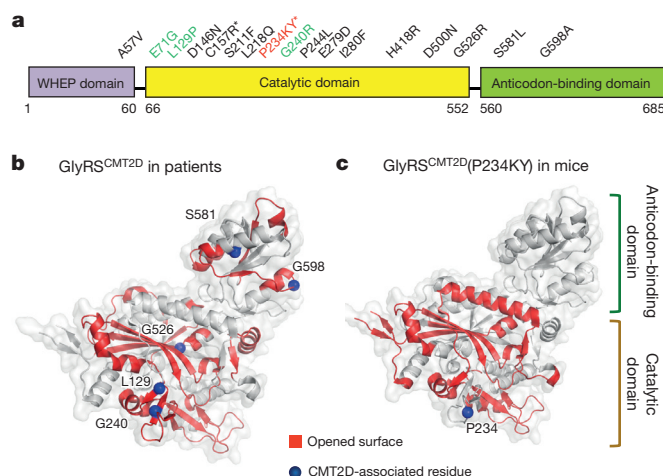


Figure 1 | Dispersed CMT2D mutations consistently cause neomorphic structural opening at the dimer interface of GlyRS. **a**, Distribution of 15 CMT2D-associated dominant mutations in the three domains of the cytosolic human GlyRS. The three strongest pathogenic mutations are highlighted in green. Two mutations identified in mice (asterisks) are labelled with their corresponding residue numbers in the human protein. **b**, Human GlyRS structure (monomeric subunit) viewed from dimer interface. Consensus opened-up areas caused by five CMT2D mutations are labelled in red¹⁷. **c**, Opened-up areas (red) by the P234KY mutation (>10% increase in deuterium incorporation relative to wild-type GlyRS).

¹Departments of Chemical Physiology and Cell and Molecular Biology, The Scripps Research Institute, La Jolla, California 92037, USA. ²Howard Hughes Medical Institute and Gene Expression Laboratory, The Salk Institute for Biological Studies, La Jolla, California 92037, USA. ³Department of Molecular Therapeutics, The Scripps Research Institute, Jupiter, Florida 33458, USA. ⁴Department of Anesthesiology, University of California San Diego, La Jolla, California 92093, USA. ⁵Molecular Neurogenetics Group, VIB Department of Molecular Genetics, University of Antwerp, BE-2610 Antwerp, Belgium. ⁶Department of Neurology, Medical University of Sofia, 1431 Sofia, Bulgaria. ⁷The Jackson Laboratory, Bar Harbor, Maine 04609, USA.

*These authors contributed equally to this work.

The new surfaces exposed by mutations in GlyRS^{CMT2D} could lead to neomorphic protein interactions. This prompted us to search for binding partners unique to GlyRS^{CMT2D}. We performed a candidate-protein screen by *in vitro* protein pull-down assays. Because motor neurons are the most frequently affected neuronal type in CMT2D^{4,18,19}, our initial screen focused on molecules that are highly expressed by motor neurons and that have been linked to motor neuron diseases or defects. We detected strong binding between the receptor Nrp1 and several GlyRS^{CMT2D} mutants including P234KY and the three (E71G, L129P and G240R) with the strongest link to CMT2D in patients¹¹ (Fig. 2a and Extended Data Fig. 2a, b). In contrast, GlyRS^{WT} did not bind Nrp1 strongly, and GlyRS^{CMT2D} failed to bind to other motor neuron proteins such as TrkB (also known as Ntrk2), Dcc, Robo1 and Unc5C (Fig. 2a and Extended Data Fig. 2a, b). To confirm that this binding specificity occurs *in vivo*, we performed immunoprecipitations using neural tissues from wild-type and P234KY-mutated CMT2D mouse littermates. Anti-Nrp1 antibodies co-precipitated

significantly more GlyRS from CMT2D mice than wild-type controls (Fig. 2b), indicating that GlyRS^{CMT2D} (P234KY) binds to Nrp1 *in vivo*. Furthermore, we also found that the GlyRS–Nrp1 interaction is significantly stronger in CMT2D patients carrying the L129P mutation than in healthy individuals (Fig. 2c). To quantify these interactions, we used biolayer interferometry and a biosensor with immobilized Nrp1 on the surface. GlyRS^{WT} binding was undetectable at 1 μ M, whereas GlyRS^{CMT2D} bound significantly more strongly, with a dissociation constant (K_d) of 29.8 ± 6.3 nM for L129P and 208.7 ± 53.8 nM for P234KY.

Next, we mapped the site where GlyRS^{CMT2D} binds to Nrp1 using pull-down assays with domain-deletion constructs. Removal of the extracellular a and c domains of Nrp1 did not alter GlyRS^{CMT2D} binding, whereas the extracellular Nrp1 b1 domain alone was sufficient to bind GlyRS^{CMT2D} (P234KY) (Fig. 2d and Extended Data Fig. 2c). Because the b1 domain is the binding site of VEGF-A₁₆₅, this finding raised the possibility that GlyRS^{CMT2D} might influence the binding of

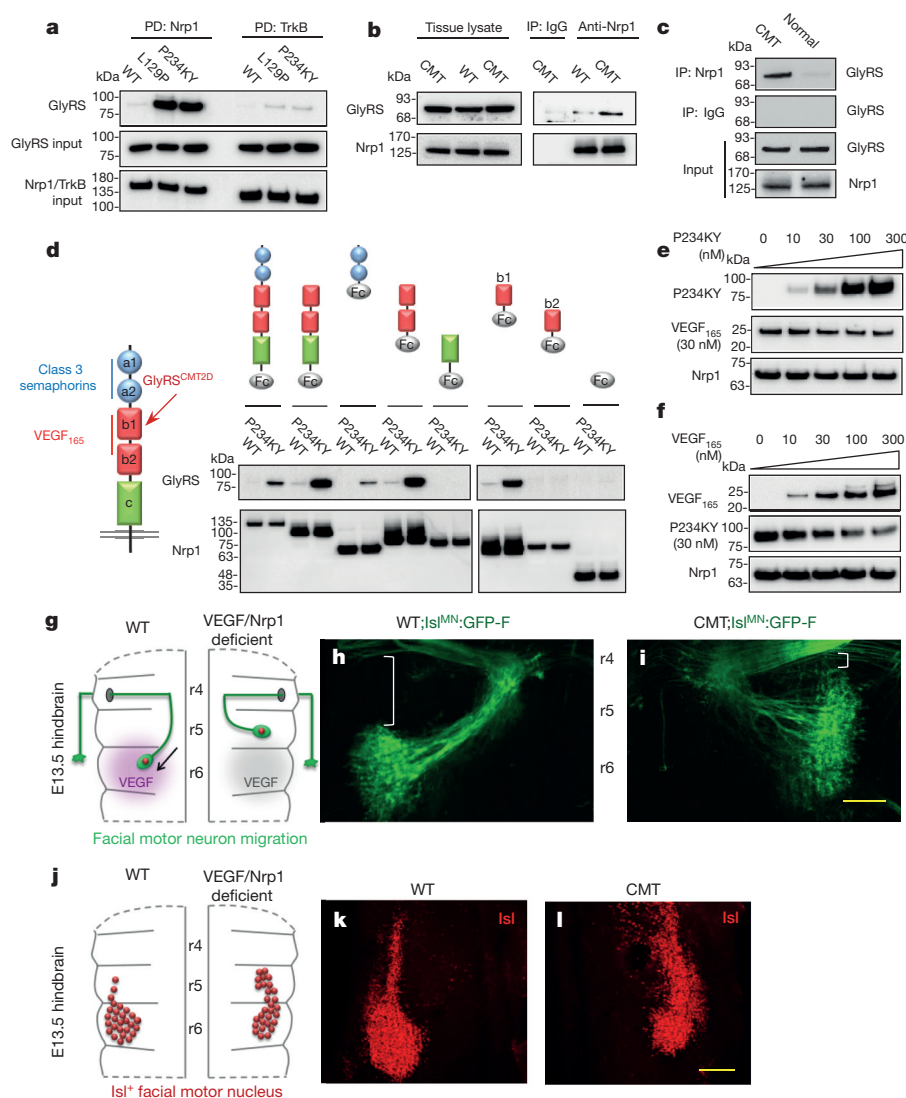


Figure 2 | GlyRS^{CMT2D} specifically binds Nrp1 and antagonizes VEGF–Nrp1 interaction. **a**, *In vitro* pull-down (PD) of GlyRS^{CMT2D} proteins by the ectodomain of Nrp1 but not TrkB. **b**, Co-immunoprecipitation (IP) to detect GlyRS–Nrp1 interaction in neural tissues of wild-type (WT) and P234KY–Gars^{CMT2D} mice (CMT). **c**, Co-immunoprecipitation to detect GlyRS–Nrp1 interaction in lymphocytes from CMT2D patients carrying the L129P mutation ($n = 5$) and from healthy individuals ($n = 3$). **d**, Domain mapping using *in vitro* pull-down identifies the b1 domain of Nrp1 as the main binding site of GlyRS^{CMT2D}. **e**, **f**, *In vitro* pull-down assay showing the competition between

GlyRS^{CMT2D} (P234KY) and VEGF-A₁₆₅ proteins for Nrp1 (b domains) binding. **g**, **j**, Schematic of facial motor neuron migration (**g**) and facial motor nucleus (**j**) in open-book preparations of wild-type (left half) and VEGF–Nrp1-deficient mouse hindbrains at E13.5 (right half). **h**, **i**, Fluorescence labelling of facial motor neuron somata and axons by Isl^{MN}:GFP-F on one side of E13.5 mouse hindbrain of open-book preparation. **k**, **l**, Immunostaining of Isl-positive facial nucleus on one side of the E13.5 mouse hindbrain of open-book preparation. Scale bars, 200 μ m.

VEGF-A₁₆₅ to this region of Nrp1. Using pull-down assays, we found that increasing concentrations of P234KY or L129P GlyRS^{CMT2D} compete with VEGF-A₁₆₅ binding to the b domains of Nrp1 (Fig. 2e and Extended Data Fig. 2d). Conversely, increasing levels of VEGF-A₁₆₅ displaced P234KY or L129P GlyRS^{CMT2D} from the b domains (Fig. 2f and Extended Data Fig. 2e).

These observations led us to focus on GlyRS protein in the extracellular environment. Recent studies of GlyRS^{WT} detected secretion from immune cells⁸, so we first examined whether GlyRS^{WT} is released by cell types relevant to the peripheral nervous system and motor function. Indeed, endogenous GlyRS^{WT} was detected in the culture media of mouse motor neuron and differentiated myotube cell lines, but not of undifferentiated myoblasts (Extended Data Fig. 3a–e). We found that secreted GlyRS^{WT} was enriched from extracellular sources using procedures that concentrate micro-vesicles (30–100 nm, ‘exosomes’) (Extended Data Fig. 4a, b). Extracellular levels of GlyRS^{WT} were diminished by application of the exosome-pathway inhibitor GW4869 and enhanced by the exosome-pathway activator monensin (Extended Data Fig. 3a–f). Next, we checked whether CMT2D-causing mutations affect the secretion of GlyRS^{CMT2D}. Our results showed that GlyRS^{CMT2D} (P234KY) was detected at levels similar to GlyRS^{WT} in the media of transfected cells (Extended Data Fig. 3g).

Nrp1 is a well-established receptor needed for motor neuron axon guidance and cell body migration^{20,21}. Our finding that GlyRS^{CMT2D} and VEGF-A₁₆₅ compete for access to Nrp1 raised the possibility that CMT2D mice may phenocopy some features of VEGF-A₁₆₅ (the mouse equivalent of human VEGF-A₁₆₅) and Nrp1 mutant mice²⁰. VEGF–Nrp1 signalling is necessary for the caudal migration of facial motor neurons from rhombomere (r)4 to r6 during embryonic development (Fig. 2g, j)²⁰. This provided us with an excellent *in vivo* assay to examine the effect of GlyRS^{CMT2D} on a well-characterized system known to depend on VEGF–Nrp1 signalling. To track facial motor neuron migration we crossed CMT2D mice to transgenic Isl^{MN}:GFP-F reporter animals that selectively label facial motor neurons^{22,23}. We found that CMT2D mutant embryos developed at a normal rate, appeared overtly normal based on their overall morphology, and that the expression levels of a variety of neuronal proteins were comparable to controls (Extended Data Fig. 5a, b). However, green fluorescent protein (GFP)-labelled facial motor neuron somata were found in ectopic anterior rhombomere locations in embryonic day (E)13.5 GlyRS^{CMT2D} mutant embryos and manifested as an elongated stream across multiple rhombomeres (Fig. 2i). In contrast, most facial cells had completed their caudal migration to r6 in littermate controls at this stage (Fig. 2h). This observation was confirmed by immunostaining with the LIM homeodomain transcription factor Isl that is selectively expressed in the nuclei of facial branchiomotor neurons (Fig. 2k, l and Extended Data Fig. 5c). This facial motor neuron migration defect closely resembles the phenotypes of Nrp1-null and VEGF-A₁₆₅-null mice, as previously reported (Fig. 2j)²⁰. Taken together, the protein-binding studies and embryological defects associated with GlyRS^{CMT2D} suggested that GlyRS^{CMT2D} inhibits VEGF–Nrp1 signalling.

VEGF signalling is thought to protect neurons from a variety of damaging insults²⁴. Intriguingly, deficient VEGF signalling leads to the selective degeneration of motor neurons in mice²⁵. To examine whether the VEGF–Nrp1 pathway is involved in motor deficits that arise from GlyRS mutations, we tested whether a genetic interaction could be detected between *Gars*^{CMT2D} and *Nrp1*. Although the data described earlier suggest that GlyRS^{CMT2D} attenuates normal Nrp1 signalling, it is still possible that GlyRS^{CMT2D} might activate aberrant signalling through Nrp1 and cause motor defects. In the first case motor phenotypes should get more severe as *Nrp1* gene dosage is reduced in CMT2D mice, and in the second case motor phenotypes should improve. To test these possibilities, we intercrossed *Gars*^{CMT2D} mice with *Nrp1* heterozygous (*Nrp1*^{+/-}) animals and characterized the motor behaviour of the single- and double-mutant offspring. At 2

weeks, when motor behavioural changes were not observed in either *Gars*^{CMT2D} or *Nrp1*^{+/-} mutant mice, 20% of the compound heterozygous *Gars*^{CMT2D}/*Nrp1*^{+/-} mutant mice had developed neuromuscular dysfunction based on a hindlimb extension test (Extended Data Fig. 6a, b). At 4 weeks, CMT2D-like symptoms, including overt neuromuscular dysfunction and an altered walking stride, become apparent in *Gars*^{CMT2D} mutants, whereas *Nrp1*^{+/-} mice appeared normal (Fig. 3a–d and Supplementary Videos 1, 2). Strikingly, by 4 weeks, 50% of the *Gars*^{CMT2D}/*Nrp1*^{+/-} mutant mice had entirely lost the ability to spread their legs and toes (Fig. 3a, b), and exhibited severely abnormal gait patterns (Fig. 3c, d and Supplementary Video 3). After postnatal week 4, *Gars*^{CMT2D}/*Nrp1*^{+/-} mutant mice began to die. Consistent with our biochemical studies showing that GlyRS^{CMT2D} binds poorly to other signalling receptors (see Fig. 2a and Extended Data Fig. 2a), intercrosses between *Gars*^{CMT2D} and *TrkB*^{+/-}, *Dcc*^{+/-}, *Robo1*^{+/-} and *Unc5C*^{+/-} heterozygous mice did not worsen the neuromuscular

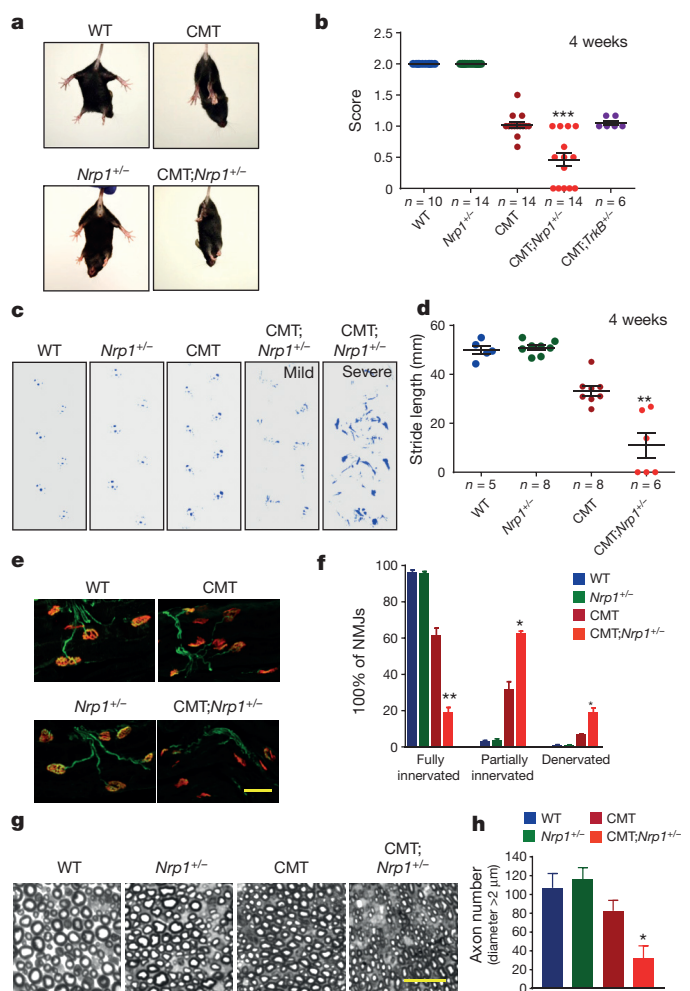


Figure 3 | *Nrp1* is a genetic modifier of CMT2D. **a, b**, Hindlimb extension test at 4 weeks. ****P* < 0.001 (Mann–Whitney test). CMT, P234KY-*Gars*^{CMT2D} mice. **c**, Hindlimb footprints of wild-type and mutant animals at 4 weeks. *Gars*^{CMT2D}/*Nrp1*^{+/-} mutant mice exhibit disrupted gait patterns of different degrees (mild, severe). Note that severe cases show inability to walk. **d**, Stride length of wild-type and mutant animals at 4 weeks. **e, f**, Neuromuscular junctions immunostaining (NMJ) in the gastrocnemius muscles of 4-week-old mice with the motor nerve terminal labelled in green and acetylcholine receptors on the muscle labelled in red. Data are presented as mean values ± standard error of the mean (s.e.m.). *n* = 3 mice per group. Scale bar, 50 μm. **g, h**, Myelinated axons from sciatic nerves of 4-week-old mice. Scale bar, 20 μm. **h**, Histogram showing the quantification of axon numbers with the diameter larger than 2 μm. *n* = 3 mice per group. **d, f, h**, **P* < 0.05, ***P* < 0.01 (*t*-test).

phenotypes in the compound heterozygotes (Fig. 3b and Extended Data Fig. 6c).

The motor defects in *Gars*^{CMT2D} and *Gars*^{CMT2D}/*Nrp1*^{+/-} mutants were accompanied by marked pathological changes in the peripheral nerves and synaptic contacts with muscle fibres. Neuromuscular junctions (NMJs) displayed a normal apposition of nerve fibres and post-synaptic acetylcholine receptors in wild-type and *Nrp1*^{+/-} animals, while partially innervated and completely denervated NMJs were present in 4-week *Gars*^{CMT2D} mutants (Fig. 3e, f). The loss of nerve terminals at NMJs was markedly increased in *Gars*^{CMT2D}/*Nrp1*^{+/-} mutants (Fig. 3e, f). Likewise, at 4 weeks of age many large-diameter axons were absent from the sciatic nerves of *Gars*^{CMT2D} mutants compared with wild-type and *Nrp1*^{+/-} littermates (Fig. 3g, h and Extended Data Fig. 7a–c). The absence of large-diameter axons was even more dramatic in 4-week *Gars*^{CMT2D}/*Nrp1*^{+/-} compound mutants, and was comparable to the extreme axonal dystrophy observed in late-stage CMT2D mutants (Fig. 3g, h and Extended Data Fig. 7d)¹⁵. These findings demonstrate that *Nrp1* is an important genetic modifier of CMT2D pathology and that GlyRS^{CMT2D} antagonizes normal *Nrp1* signalling rather than activating aberrant signalling.

These findings prompted us to test whether VEGF overexpression could counteract GlyRS^{CMT2D} and help to slow the loss of motor function in CMT2D mice. A lentiviral vector encoding either VEGF-A₁₆₅ or GFP was injected bilaterally into the hindlimb muscles of *Gars*^{CMT2D} mutant mice at postnatal day 5, before the onset of overt motor defects (Fig. 4a and Extended Data Fig. 8). By 4 weeks of age, we began to observe a reduction of limb strength in the control GFP-treated *Gars*^{CMT2D} animals using an inclined plane test (Fig. 4b). However, VEGF-A₁₆₅-treated animals retained greater neuromuscular capacity with significantly higher scores (Fig. 4b). By 7 weeks, *Gars*^{CMT2D} animals exhibited a disrupted gait pattern with shortened hindlimb stride length, while VEGF-A₁₆₅-treated animals maintained a significantly longer walking stride (Fig. 4c). Likewise, VEGF treatment significantly improved the motor performance of *Gars*^{CMT2D} mutants in the rotarod test (Fig. 4d). To minimize the possible influence from the natural variation in disease progression among individual animals, we compared the effect of VEGF treatment to GFP controls by separately treating each hindlimb from the same animal.

Lentiviral vectors encoding VEGF-A₁₆₅ or GFP were injected unilaterally into each hindlimb of the same *Gars*^{CMT2D} mutant mouse at postnatal day 5 (Extended Data Fig. 9a). At 5 weeks, GFP-treated hindlimbs developed severe muscle weakness, largely losing their ability to extend. In contrast, the contralateral hindlimbs treated with VEGF-A₁₆₅ retained significant function (Extended Data Fig. 9b–e). These results suggest that VEGF treatment significantly ameliorates the loss of motor function in CMT2D mice.

A number of neurotrophic factors have been tested as broad-spectrum strategies to enhance neuronal survival and treat motor diseases^{26–30}. This raised the possibility that VEGF might slow the progression of CMT2D pathology by functioning as a generic trophic factor rather than as a specific agent to restore normal VEGF–Nrp1 signalling. Therefore, we tested the effects of lentivirus-mediated expression of GDNF, a potent neurotrophin that has been used to enhance motor function and survival in mouse models of amyotrophic lateral sclerosis^{26,27}. Unlike VEGF-A₁₆₅, we found that GDNF failed to slow disease progression in *Gars*^{CMT2D} mice (Fig. 4b, d and Extended Data Fig. 9f). Next, we tested whether VEGF-mediated motor sparing is dependent upon *Nrp1* binding, by exploiting the binding specificity of VEGF protein isoforms. VEGF-A₁₂₁ has overlapping functions with VEGF-A₁₆₅ but lacks high-affinity *Nrp1* binding²⁴. We found that VEGF-A₁₂₁ treatment failed to ameliorate the loss of motor function in CMT2D animals (Fig. 4b, d and Extended Data Fig. 9g). These data support a model in which VEGF treatment helps to guard against the motor loss arising from the aberrant activity of GlyRS^{CMT2D} by restoring VEGF–Nrp1 signalling.

Our results identify the *Nrp1* gene as an important genetic modifier for CMT2D, and link the selective pathology of this disease to the neomorphic binding of GlyRS^{CMT2D} to the receptor *Nrp1*. Although we found the same neomorphic binding to NRP1 in several human GlyRS^{CMT2D} mutants that are strongly disease associative (Fig. 2a and Extended Data Fig. 2a, b), we cannot rule out the possibility that some GlyRS^{CMT2D} mutants may interact with other extracellular and/or intracellular targets. Nevertheless, our findings strongly suggest that the VEGF–NRP1 pathway is an actionable target for treating CMT2D (Extended Data Fig. 10). While the exact role of VEGF in the motor system remains poorly defined, VEGF-deficient mice selectively develop symptoms of motor neuron disease over time²⁵. The direct antagonism of VEGF–Nrp1 signalling by GlyRS^{CMT2D} found here further indicates that deficient VEGF signalling may represent a common pathogenic pathway that is susceptible to abnormal activity in other motor neuron diseases. A broad implication from this work is that the molecular basis of selective neuronal vulnerability in neurodegenerative diseases may arise from the neomorphic activity of misfolded proteins interacting with susceptible signalling targets in specific cell types. This conceptual framework may be applied for identifying additional druggable targets to treat neurodegenerative diseases, including other forms of CMT.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 7 August 2014; accepted 21 August 2015.

Published online 21 October 2015.

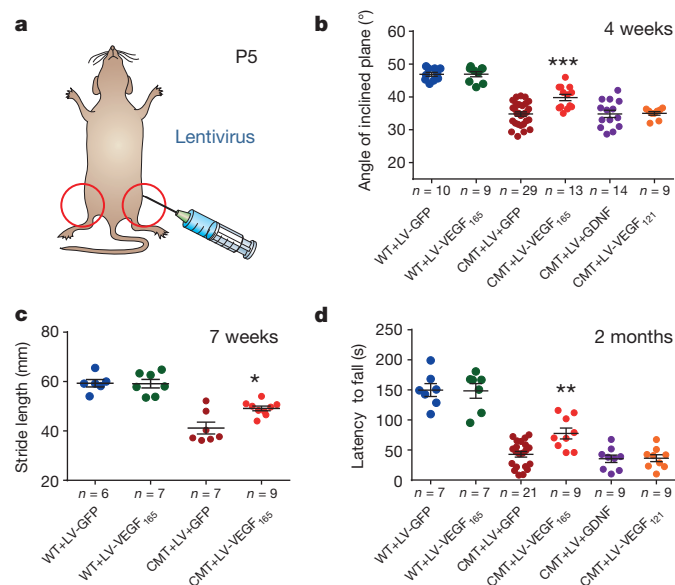


Figure 4 | VEGF treatment improves motor function in CMT2D mice.

a, Diagram showing bilateral intramuscular injection of lentivirus (LV) into mouse hindlimbs at postnatal day (P)5. **b**, Inclined plane test of 4-week-old animals. **c**, Walking strides of 7-week-old animals. **d**, Rotarod test of 2-month-old animals. **P* < 0.05, ***P* < 0.01, ****P* < 0.001 (*t*-test).

- Saxena, S. & Caroni, P. Selective neuronal vulnerability in neurodegenerative diseases: from stressor thresholds to degeneration. *Neuron* **71**, 35–48 (2011).
- Skre, H. Genetic and clinical aspects of Charcot-Marie-Tooth's disease. *Clin. Genet.* **6**, 98–118 (1974).
- Patzkó, A. & Shy, M. E. Update on Charcot-Marie-Tooth disease. *Curr. Neurol. Neurosci. Rep.* **11**, 78–88 (2011).
- Antonellis, A. et al. Glycyl tRNA synthetase mutations in Charcot-Marie-Tooth disease type 2D and distal spinal muscular atrophy type V. *Am. J. Hum. Genet.* **72**, 1293–1299 (2003).
- Wakasugi, K. & Schimmel, P. Two distinct cytokines released from a human aminoacyl-tRNA synthetase. *Science* **284**, 147–151 (1999).
- Yao, P. & Fox, P. L. Aminoacyl-tRNA synthetases in medicine and disease. *EMBO Mol. Med.* **5**, 332–343 (2013).

7. Williams, T. F., Mirando, A. C., Wilkinson, B., Francklyn, C. S. & Lounsbury, K. M. Secreted Threonyl-tRNA synthetase stimulates endothelial cell migration and angiogenesis. *Sci. Rep.* **3**, 1317 (2013).
8. Park, M. C. *et al.* Secreted human glycyl-tRNA synthetase implicated in defense against ERK-activated tumorigenesis. *Proc. Natl Acad. Sci. USA* **109**, E640–E647 (2012).
9. Guo, M., Yang, X. L. & Schimmel, P. New functions of aminoacyl-tRNA synthetases beyond translation. *Nature Rev. Mol. Cell Biol.* **11**, 668–674 (2010).
10. Lee, H. J. *et al.* Two novel mutations of GARS in Korean families with distal hereditary motor neuropathy type V. *J. Peripher. Nerv. Syst.* **17**, 418–421 (2012).
11. Motley, W. W., Talbot, K. & Fischbeck, K. H. GARS axonopathy: not every neuron's cup of tRNA. *Trends Neurosci.* **33**, 59–66 (2010).
12. Kawakami, N. *et al.* A novel mutation in glycyl-tRNA synthetase caused Charcot-Marie-Tooth disease type 2D with facial and respiratory muscle involvement. *Rinsho Shinkeigaku* **54**, 911–915 (2014).
13. Sun, A. *et al.* A novel mutation of the glycyl-tRNA synthetase (GARS) gene associated with Charcot-Marie-Tooth type 2D in a Chinese family. *Neurol. Res.* **37**, 782–787 (2015).
14. Achilli, F. *et al.* An ENU-induced mutation in mouse glycyl-tRNA synthetase (GARS) causes peripheral sensory and motor phenotypes creating a model of Charcot-Marie-Tooth type 2D peripheral neuropathy. *Dis. Model. Mech.* **2**, 359–373 (2009).
15. Seburn, K. L., Nangle, L. A., Cox, G. A., Schimmel, P. & Burgess, R. W. An active dominant mutation of glycyl-tRNA synthetase causes neuropathy in a Charcot-Marie-Tooth 2D mouse model. *Neuron* **51**, 715–726 (2006).
16. Motley, W. W. *et al.* Charcot-Marie-Tooth-linked mutant GARS is toxic to peripheral neurons independent of wild-type GARS levels. *PLoS Genet.* **7**, e1002399 (2011).
17. He, W. *et al.* Dispersed disease-causing neomorphic mutations on a single protein promote the same localized conformational opening. *Proc. Natl Acad. Sci. USA* **108**, 12307–12312 (2011).
18. Del Bo, R. *et al.* Coexistence of CMT-2D and distal SMA-V phenotypes in an Italian family with a GARS gene mutation. *Neurology* **66**, 752–754 (2006).
19. Dubourg, O. *et al.* The G526R glycyl-tRNA synthetase gene mutation in distal hereditary motor neuropathy type V. *Neurology* **66**, 1721–1726 (2006).
20. Schwarz, Q. *et al.* Vascular endothelial growth factor controls neuronal migration and cooperates with Semaphorin 3A to pattern distinct compartments of the facial nerve. *Genes Dev.* **18**, 2822–2834 (2004).
21. Huber, A. B. *et al.* Distinct roles for secreted semaphorin signaling in spinal motor axon guidance. *Neuron* **48**, 949–964 (2005).
22. Song, M. R. *et al.* T-Box transcription factor Tbx20 regulates a genetic program for cranial motor neuron cell body migration. *Development* **133**, 4945–4955 (2006).
23. Lewcock, J. W., Genoud, N., Lettieri, K. & Pfaff, S. L. The ubiquitin ligase Phr1 regulates axon outgrowth through modulation of microtubule dynamics. *Neuron* **56**, 604–620 (2007).
24. Mackenzie, F. & Ruhrberg, C. Diverse roles for VEGF-A in the nervous system. *Development* **139**, 1371–1380 (2012).
25. Oosthuysen, B. *et al.* Deletion of the hypoxia-response element in the vascular endothelial growth factor promoter causes motor neuron degeneration. *Nature Genet.* **28**, 131–138 (2001).
26. Wang, L. J. *et al.* Neuroprotective effects of glial cell line-derived neurotrophic factor mediated by an adeno-associated virus vector in a transgenic animal model of amyotrophic lateral sclerosis. *J. Neurosci.* **22**, 6920–6928 (2002).
27. Acsadi, G. *et al.* Increased survival and function of SOD1 mice after glial cell-derived neurotrophic factor gene therapy. *Hum. Gene Ther.* **13**, 1047–1059 (2002).
28. Kaspar, B. K., Lladó, J., Sherkat, N., Rothstein, J. D. & Gage, F. H. Retrograde viral delivery of IGF-1 prolongs survival in a mouse ALS model. *Science* **301**, 839–842 (2003).
29. Nayak, M. S., Kim, Y. S., Goldman, M., Keirstead, H. S. & Kerr, D. A. Cellular therapies in motor neuron diseases. *Biochim. Biophys. Acta* **1762**, 1128–1138 (2006).
30. Azzouz, M. *et al.* VEGF delivery with retrogradely transported lentivector prolongs survival in a mouse ALS model. *Nature* **429**, 413–417 (2004).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank P. Schimmel, J. Dolkas, C. Farrokhi, L. Lisowski, C. Ly, S. Chalasani, T. Liu, Q. Hu, P. Li, N. Sheng and members of The Scripps Laboratories of tRNA Synthetase Research and the Pfaff laboratory for advice and discussions on the experiments and manuscript. We are grateful to S. Ackerman, K.-F. Lee, D. O'leary and D. Ginty for providing mouse lines. W.H., H.Z. and N.W. were supported by fellowships from the National Foundation for Cancer Research. G.B. was supported by the Pioneer fund and Howard Hughes Medical Institute. S.L.P. is a Howard Hughes Medical Institute investigator and a Benjamin H. Lewis chair in Neuroscience. This research is supported by grants from the US National Institutes of Health (R01GM088278, R21NS084254 and R01NS054154), The Marshall Heritage Foundation and the Sol Goldman Trust, and by aTyr Pharma through an agreement with The Scripps Research Institute.

Author Contributions W.H., G.B., S.L.P. and X.-L.Y. designed the study, analysed the data, and prepared the manuscript. W.H. carried out molecular cloning, binding analyses, secretion studies, and other biochemical experiments. G.B. performed the mouse genetics, viral injections, behaviour testing, and histology experiments. H.Z. carried out Nrp1 domain mapping, GlyRS-VEGF competition, and additional pull-down assays, and contributed to study design and figure preparation. N.W. performed co-immunoprecipitation to detect aberrant GlyRS-Nrp1 interaction using CMT2D patient samples. A.J. and V.G. provided transformed lymphocytes samples from CMT2D patients. J.L. and P.R.G. performed the hydrogen-deuterium exchange analysis. N.M.W. and K.L. assisted with mouse studies. Y.S. and C.D.D. contributed to biochemical experiments. H.L. and V.S. contributed to histology experiments. R.W.B. provided mice, technical support, and scientific advice.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to X.-L.Y. (xlyang@scripps.edu) or S.L.P. (pfaff@salk.edu)

METHODS

Mice. The following strains of mice were used in this study: wild-type C57BL/6J (JAX), P234KY-CMT2D mutant (background includes a mix of C57BL/6, CB6 and CAST)¹⁴, Tg (Hb9-GFP)³¹, Tg (Isl^{MIN}:GFP-F)²³, Nrpl mutant³², TrkB mutant³³, Robo1 mutant³⁴, Dcc mutant³⁵, and Unc5C mutant³⁶. Both male and female mice were used in this study. All experiments were done in accordance with Institutional Animal Care and Use Committee animal protocols and BSL2+ safety protocols, on animals housed in groups on a 12-h light–dark cycle.

Recombinant GlyRS expression and purification. Carboxy-terminal His-tagged human GlyRS^{WT} and GlyRS^{CMT2D} proteins were individually cloned into pET21b vector (Novagen) and expressed in *Escherichia coli* BL21 (DE3) host cells at 25 °C. The proteins were purified by Ni-NTA agarose affinity column followed by ion-exchange monoQ column and size-exclusion column Superdex 200 (GE Healthcare). To prepare non-tagged human GlyRS^{WT} and GlyRS^{CMT2D} proteins, the GARS gene was fused with an amino-terminal His-SUMO tag, cloned into pET28a vector (Novagen), and expressed as His-SUMO-GlyRS fusion proteins in *Escherichia coli* BL21 (DE3) cells. The fusion proteins were purified with a Ni-NTA agarose affinity column, and then subjected to homemade Ulp1 protease to remove the His-SUMO tag. The non-tagged GlyRS proteins were separated from the tag by flowing through the Ni-NTA column again.

Hydrogen–deuterium exchange analysis. Solution-phase amide hydrogen–deuterium exchange (HDX) was performed with a fully automated system as described previously³⁷. Briefly, 4 µl of His-tagged GlyRS^{P234KY} or GlyRS^{WT} was diluted to 20 µl with D₂O-containing HDX buffer to a final concentration of 10 µM, and incubated at 4 °C for 10, 30, 60, 900, and 3,600 s. Following on-exchange, unwanted back exchange was minimized by adding 30 µl of 1% TFA in 5 M urea to denature the protein (held at 1 °C). Samples were then passed across an immobilized pepsin column (prepared in house) at 50 µl min^{−1} (0.1% TFA, 15 °C), and the resulting peptides were trapped onto a C8 trap cartridge (Thermo Fisher, Hypersil Gold). Peptides were eluted across a 1 mm × 50 mm C18 high-performance liquid chromatography (HPLC) column (Hypersil Gold, Thermo Fisher) with a 4–40% CH₃CN gradient and 0.3% formic acid over 5 min at 2 °C, and electrosprayed directly into an Orbitrap mass spectrometer (LTQ Orbitrap with ETD; Thermo Fisher). Data were processed with in-house software³⁸ and visualized with PyMOL (DeLano Scientific). The difference in HDX between GlyRS^{P234KY} and GlyRS^{WT} was calculated by subtracting the average percentage deuterium uptake for GlyRS^{P234KY} from that for GlyRS^{WT} after 10, 30, 60, 300, 900 and 3,600 s of on-exchange. Please note that we cannot directly compare the numbers obtained here for GlyRS^{P234KY} with those for other GlyRS^{CMT2D} mutants from the previous study¹⁷, because this and the previous analysis were carried out in two different laboratories with different instruments and experimental procedures.

Detection of GlyRS proteins in cell cultures. NSC-34 motor neuron cells (Cellutions Biosystems) and C2C12 mouse adherent myoblasts (from A. Patapoutian's laboratory at The Scripps Research Institute) were maintained in DMEM supplemented with 10% heat-inactivated fetal bovine serum (FBS) and 1% penicillin (Life Technologies) at 37 °C in a humidified incubator containing 5% CO₂. These cell lines had not been recently authenticated and tested for mycoplasma contamination. Myogenic differentiation of C2C12 myoblasts was induced by substituting the FBS with 2% horse serum. The cells were further cultured in Opti-MEM Reduced Serum Medium (Life Technologies) for 16 h. Brefeldin A (Cell Signaling Technology), GW4869 (Sigma-Aldrich), or monensin (eBioscience) was added to the cell medium for 2 h before the cells and the medium were separated for western blot analysis. After removal of cell debris by spinning the medium at 300g for 10 min, the supernatant was concentrated using Amicon Ultra-4 Centrifugal filter (Millipore). Cells were lysed using cell lysis buffer (ATCC) with added protease inhibitor cocktail (Roche). The following antibodies were used for western blot analysis: mouse anti-GlyRS (H00002617-B01P, Abnova; 1:1,000), rabbit anti-GAPDH (#3683, Cell Signaling Technology; 1:1,000), rabbit anti-vWF (sc-14014, Santa Cruz; 1:50), and rabbit anti-TSG101 (MAB649, Millipore; 1:1,000). To study the effect of CMT2D-causing mutation on GlyRS secretion, constructs overexpressing V5-tagged GlyRS^{P234KY} or GlyRS^{WT} were transfected into COS7 cells using lipofectamine 2000 (Invitrogen). The expression and secretion of GlyRS proteins were detected by western blot analysis using anti-V5 antibody (R960-25, Invitrogen; 1:5,000).

Exosome purification and analysis. The general idea of exosome purification by differential centrifugation is depicted in Extended Data Fig. 3a. Supernatants from NSC-34 cell medium were subjected to successive centrifugation steps at 4 °C: (1) 200g for 10 min to eliminate floating cells; (2) 2,000g for 10 min to discard large dead cells; (3) 10,000g for 1 h to remove cell debris and cellular organelles such as mitochondria and lysosomes. At each step, the pellet was thrown away and the supernatant was used for the following step. The final supernatant was centrifuged at 100,000g at 4 °C for 4 h to pellet micro-vesicles that are commonly known as

'exosomes'. The final supernatant and the exosome fraction were analysed by western blot analysis using the antibodies specified earlier and rabbit anti-Bip antibody (#3183S, Cell Signaling; 1:1,000).

In vitro pull-down assay. Recombinant rat Nrpl-Fc, mouse TrkB-Fc, mouse Dcc-Fc, rat Robo1-Fc and human Unc5c-Fc extracellular domain-Fc chimaeras (R&D systems) were bound to Protein G beads. Purified non-tagged GlyRS^{WT} and GlyRS^{CMT2D} proteins were individually added to the receptor-immobilized beads and incubated for 1 h at 4 °C. After removal of unbound GlyRS proteins, SDS-loading buffer was directly added to the beads to elute the receptor and its bound GlyRS. The amount of GlyRS bound to receptors was analysed by western blot analysis using mouse anti-GlyRS antibodies (H00002617-B01P, Abnova; 1:1,000).

Co-immunoprecipitation and western blot analysis using mouse tissues. The interaction between endogenous GlyRS and Nrpl proteins was detected by co-immunoprecipitation. Adult mouse neural samples were lysed using RIPA buffer (Cell Signaling) containing 20 mM Tris-HCl (pH 7.5), 150 mM NaCl, 1 mM Na₂EDTA, 1 mM EGTA, 1% NP-40, 1% sodium deoxycholate, 2.5 mM sodium pyrophosphate, 1 mM β-glycerophosphate, 1 mM Na₃VO₄, and 1 µg ml^{−1} leupeptin. Immunoprecipitation was performed with rabbit anti-Nrpl antibody (NBP1 40666, Novus Biologicals; 1:100) and the precipitates were subjected to western blot analysis using mouse anti-GlyRS antibody (H00002617-B01P, Abnova; 1:1,000).

Western blot was performed to analyse the expression levels of various neuronal proteins in E12.5 wild-type and CMT2D mutant neural tissues. The following primary antibodies were used: mouse anti-GlyRS (H00002617-B01P, Abnova; 1:1,000), rabbit anti-Nrpl (#3725S, Cell Signaling; 1:1,000), rabbit anti-VEGFR1 (#36-1100, Life Technologies; 1:1,000), rabbit anti-VEGFR2 (#2479S, Cell Signaling; 1:1,000), mouse anti-β-actin (#3700, Cell Signaling; 1:1,000), mouse anti-Dcc (AF5, Abcam; 1:100), rabbit anti-Robo1 (NB100-60458, Novus Biologicals; 1:2,000), mouse anti-NF (2H3, DSHB; 1:100), mouse anti-MAP2 (MAB364, Millipore; 1:500), rabbit anti-GAP43 (AB5220, Millipore; 1:1,000), and rabbit anti-β-catenin (#9587, Cell Signaling; 1:1,000).

Co-immunoprecipitation using CMT2D patient samples. Peripheral blood was drawn from CMT2D patients carrying the L129P mutation and control individuals after obtaining their written informed consent. The study complies with the ethical guidelines of the Medical University of Sofia, Bulgaria and University of Antwerp, Belgium. Lymphocytes were isolated on a Ficoll–Paque gradient, transfected with Epstein–Barr virus and incubated at 37 °C for 2 h. After centrifugation, cells were re-suspended in 4 ml RPMI complete medium (Life Technologies) supplemented with 1% phytohaemagglutinin. Cells were seeded on a 24-well plate and incubated at 37 °C, 6% CO₂ for a minimum of 3 days. After establishment, cell lines were cultivated in RPMI1640 medium containing 15% fetal calf serum, 1% sodium pyruvate, 1% 200 M L-glutamine and 2% penicillin/streptomycin. The harvested lymphoblastoid cells were lysed using RIPA buffer (Cell Signaling). Immunoprecipitation was performed with rabbit anti-Nrpl antibody (Novus Biologicals) and rabbit anti-IgG (#2729, Cell Signaling) and the pull-down samples were subjected to western blot analysis using rabbit anti-GlyRS antibody (sc-98614, Santa Cruz Biotechnology; 1:500).

Mapping of GlyRS^{CMT2D} interaction domain on Nrpl. The variants of Nrpl extracellular domain (ECD) include intact ECD (residues Arg 23–Asp 840), b1b2c domain (residues Phe 273–Asp 840), a1a2 domain (residues Arg 23–Asp 272), b1b2 domain (residues Phe 273–Phe 643), c domain (residues Thr 589–Asp 840), b1 domain (residues Phe 273–Asp 428) and b2 domain (residues Lys 425–Phe 643). These variants were designed as chimaera proteins containing a 17-residue secretion signal peptide from myeloid cell surface antigen CD33 (gp67) at the N terminus and a human IgG Fc domain at C terminus, and were expressed using pcDNA6.0/V5-His-A vector (Life Technologies). For each Nrpl variants, 3 µg plasmids were transfected using Lipofectamine 2000 (Life Technologies) into human HEK293 cells in a 6-well plate. Twenty hours after transfection, MEM media containing secreted Nrpl variants were collected and incubated with 30 µl Protein A resins. The Nrpl-bound resins were divided equally into two 1.5 ml Eppendorf tubes and incubated with 5 µg of recombinant GlyRS^{CMT2D} or GlyRS^{WT} in 1 ml of Washing Buffer (PBS, 5 mM β-ME, 0.2% BSA and 0.05% Triton X-100) for 1 h. Resins were then washed three times with the Washing Buffer and once with PBS. The bound proteins were eluted with 30 µl of SDS-PAGE sample buffer and subjected to western blotting analysis using mouse anti-GlyRS (H00002617-B01P, Abnova; 1:1,000) and rabbit anti-His antibodies (RHIS-45P-Z, ICL Lab; 1:10,000) to detect GlyRS and the Nrpl variants, respectively.

The b1 (residues Phe 273–Asp 428), b2 (residues Lys 425–Glu 586), and b1b2 domain (residues Phe 273–Glu 586) of Nrpl fused with an N-terminal glutathione S-transferase (GST) tag was cloned into the pET28a vector (Novagen), expressed in *E. coli* BL21(DE3) cells and purified with GST resin (Qiagen). GST or GST–Nrpl fusion proteins was incubated with 20 µl GST resin and then bind with non-tagged wild-type or P234KY GlyRS in 1 ml of Washing Buffer (1×PBS,

5 mM BME, 0.2% BSA and 0.05% Triton X-100) for 1 h. GST resins were washed three times with Washing Buffer and once more with PBS. The bound proteins were eluted with SDS–PAGE sample buffer, and subjected to western blotting analysis.

Competition assay between VEGF-A₁₆₅ and GlyRS^{CMT2D} for Nr1 binding. In each experiment, 5 µg of GST–b1b2 protein was bound with 15 µl of GST resin in 1 ml Washing Buffer on ice. The competition was tested in both directions. In one direction, 30 nM of P234KY GlyRS^{CMT2D} was added to GST–b1b2 with an increasing concentration of human VEGF-A₁₆₅ (IBL); in the opposite direction, 30 nM of VEGF-A₁₆₅ was added with an increasing concentration of P234KY GlyRS^{CMT2D}. After the resins were washed three times with the Washing Buffer and once with PBS, proteins were eluted with SDS–PAGE sample buffer, and analysed by western blot using rabbit anti-VEGF-A (ABS82, Millipore; 1:2,000), mouse anti-GlyRS (H00002617-B01P, Abnova; 1:1,000) and rabbit anti-GST (#2622, Cell Signaling; 1:1,000) antibodies.

Immunostaining and imaging. Immunostaining of NMJs was performed as described¹⁴. Cocktails of the following primary antibodies were used to visualize nerves: rabbit anti-NF (AB1991, Millipore; 1:1,000), rabbit anti-synaptophysin (A0010, Dako; 1:2,000), and mouse anti-SV2 (DSHB; 1:1,000). Secondary antibodies were Alexa-488 or -647 conjugated (Molecular Probes/Invitrogen; 1:1,000). Tetramethylrhodamine-conjugated α -bungarotoxin (T-1175, Molecular Probes/Invitrogen; 1:1,000) was used to visualize acetylcholine receptors (AChRs) on muscles. The occupancy of NMJs is measured by examining the overlap of the motor nerve terminal (green) with AChRs on the muscle (red). At least 40 randomly selected NMJs were examined from each of three mutant and three control mice. The flat-mount preparations of hindbrains were performed as previously described²³. Rabbit anti-Isl1/2 (ref. 39) was used to label facial motor nuclei by whole-mount immunostaining. The distance between the facial motor nucleus and trigeminal nucleus was measured for each embryo. Each distance was further normalized to the relative distance of the wild-type facial motor nucleus. Rabbit anti-VEGF (ab52917, Abcam; 1:200) was used to determine the expression of VEGF in muscle fibres.

Bright-field and fluorescence images of whole embryos were obtained using a 0.8× objective on Zeiss Lumar.V12 fluorescence stereomicroscope. Confocal images were obtained using 10× and 20× objectives on Olympus Fluoview 1000 confocal microscope.

Hindlimb extension test. Mice were suspended by the tail and the extent of hindlimb extension was observed over 10 s. A score of 2 corresponded to a normal extension reflex in hindlimbs with splaying of toes. A score of 1 corresponded to clenching of hindlimbs to the body with partial splaying of toes. A score of 0 corresponded to clasping hindlimbs with curled toes. Three tests were performed for each mouse with 5 s intervals. A score of 1.5 or 0.5 corresponded to behaviours between 2 and 1, or between 1 and 0, respectively.

Footprint test. Blue ink was applied to the hind paws of each mouse and the animal was placed in a narrow alley (9 × 80 × 25 cm) with the floor covered with white paper. A home cage was placed at the end of the alley for the animal to walk to while leaving its footprints on the paper. Stride length was assessed by measuring the average distances of at least three consecutive steps on each side.

Inclined plane test. Hindlimb strength was assessed at postnatal 4 weeks using the inclined plane test. Briefly, animals were placed on an inclined plane, and the angle of incline was gradually increased starting from 15°. The maximum angle at which the animal could maintain its position for 5 s constituted the inclined plane score. The test was performed three times for each mouse.

Rotarod test. Motor coordination was assessed with a rotarod apparatus (Economex, Columbus Instruments). The mice were first placed on the stationary rod (0 r.p.m.) to acclimate them to the apparatus, followed by a trial at a rotation speed of 1 r.p.m. for 3 min or until a fall occurred. For testing, the rotation of the rotarod was accelerated from 0 r.p.m. with an accelerating rate (0.1 r.p.m. min⁻¹). The latency of each mouse to fall was monitored for three consecutive trials and the intra-trial interval for each animal was about 20 min. The average time of three trials was used as a measure of motor performance.

Virus preparation and injection. The complementary DNAs encoding GDNF, VEGF-A₁₂₁ or VEGF-A₁₆₅ were cloned into lentiviral vector (p156RRLsin PPTCMVGFPRE) between BamHI and SalI sites. All lentiviruses were produced by the GT3 core facility at the Salk Institute with a titre of 1×10^{12} – 2×10^{13} genome copies per ml. Injections were performed at P5 (\pm 1 day) after anaesthetizing pups on ice. Multiple injections ($n \geq 8$) of virus (5 µl for each limb) into a variety of hindlimb muscles were performed with a Hamilton syringe. On the basis of the expression pattern of GFP reporter, we found that the lentiviruses mainly infect muscle fibres.

Nerve histology and imaging. Mouse sciatic nerves were dissected and fixed in 2.5% glutaraldehyde in 0.1 M phosphate buffer. Nerve samples were then osmicated, dehydrated and embedded in araldite resin. Transverse nerve sections (1 µm) were cut on a Leica RM2065 microtome and stained with methylene blue Azure II. Images were collected on a Leica DMR microscope or an Olympus BX61 microscope. Axon numbers were determined from two non-overlapping fields (50 × 50 µm) from each of three mutant and three control nerve samples. Axon diameters were measured by Image J.

Statistics. All graphs and data generated in this study were analysed using GraphPad Prism 6.0 Software (MacKiev) or Excel (Microsoft). Two-tailed unpaired *t*-tests with Welch's correction using parametric distribution, two-tailed Mann–Whitney test using unparametric distribution, or two-tailed paired Wilcoxon test using unparametric distribution were performed to measure differences from at least three independent biological replicates. $P < 0.05$ was considered significant. These tests do not require similar variance of the data between the groups that are being statistically compared. The normality of the data was determined by D'Agostino–Pearson omnibus test and Kolmogorov–Smirnov test. No statistical methods were used to predetermine sample sizes, but our sample sizes are similar to those generally used in the field.

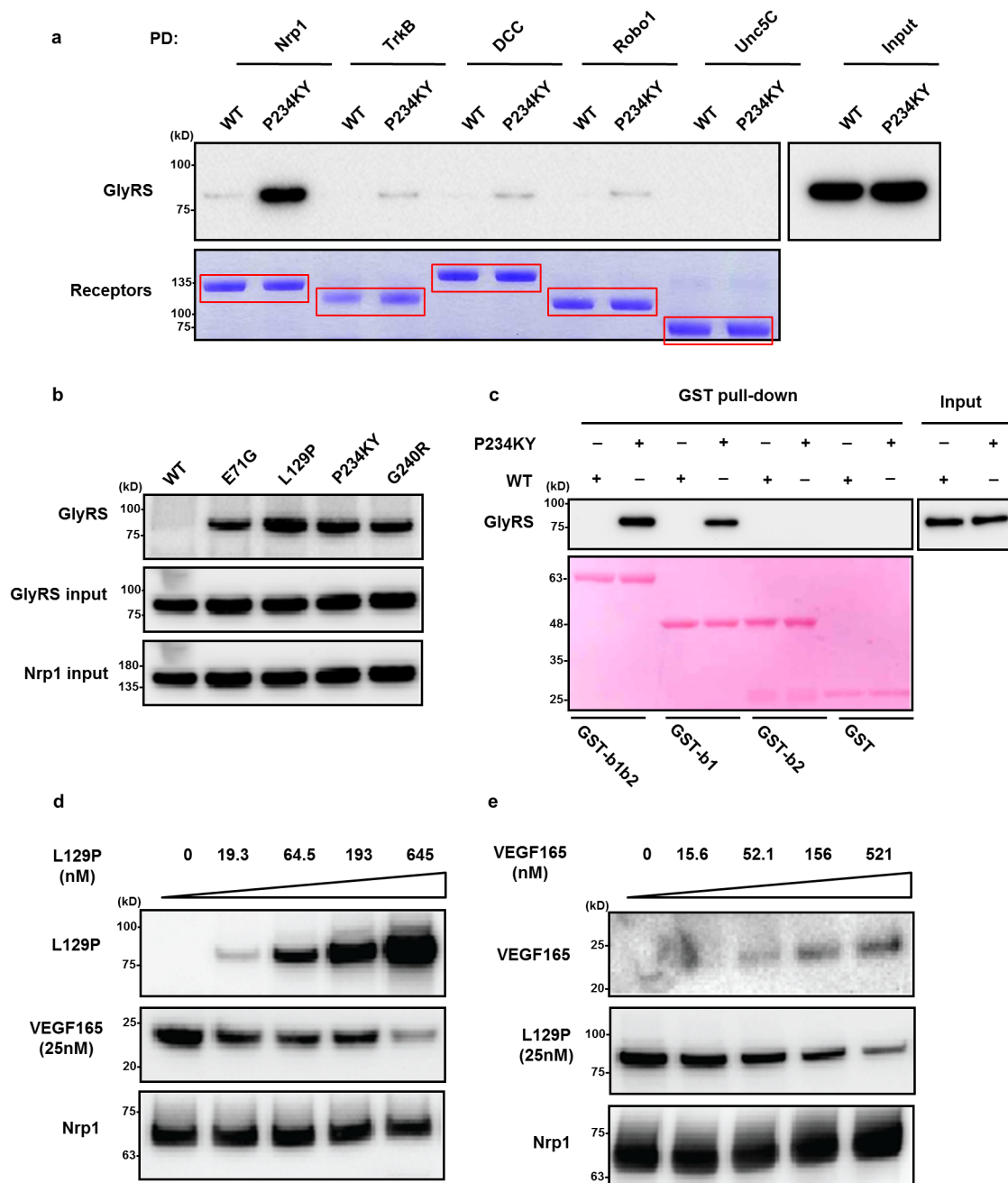
For all animal studies, analyses were performed on approximately equal numbers of male and female mice selected randomly from populations, and no sex-specific differences in the disease progression were identified. All behavioural experiments were performed in double-blind fashion, and stressed animals were excluded from the analysis.

- Lee, S. K., Jurata, L. W., Funahashi, J., Ruiz, E. C. & Pfaff, S. L. Analysis of embryonic motoneuron gene regulation: derepression of general activators function in concert with enhancer factors. *Development* **131**, 3295–3306 (2004).
- Gu, C. *et al.* Neurofilin-1 conveys semaphorin and VEGF signaling during neural and cardiovascular development. *Dev. Cell* **5**, 45–57 (2003).
- Xu, B. *et al.* Cortical degeneration in the absence of neurotrophin signaling: dendritic retraction and neuronal loss after removal of the receptor TrkB. *Neuron* **26**, 233–245 (2000).
- Ma, L. & Tessier-Lavigne, M. Dual branch-promoting and branch-repelling actions of Slit/Robo signaling on peripheral and central branches of developing sensory axons. *J. Neurosci.* **27**, 6843–6851 (2007).
- Fazeli, A. *et al.* Phenotype of mice lacking functional *Deleted in colorectal cancer* (*Dcc*) gene. *Nature* **386**, 796–804 (1997).
- Burgess, R. W., Jucius, T. J. & Ackerman, S. L. Motor axon guidance of the mammalian trochlear and phrenic nerves: dependence on the netrin receptor Unc5c and modifier loci. *J. Neurosci.* **26**, 5756–5766 (2006).
- Chalmers, M. J. *et al.* Probing protein ligand interactions by automated hydrogen/deuterium exchange mass spectrometry. *Anal. Chem.* **78**, 1005–1014 (2006).
- Pascal, B. D. *et al.* HDX workbench: software for the analysis of H/D exchange MS data. *J. Am. Soc. Mass Spectrom.* **23**, 1512–1521 (2012).
- Ericson, J., Thor, S., Edlund, T., Jessell, T. M. & Yamada, T. Early stages of motor neuron differentiation revealed by expression of homeobox gene *Isl-1*. *Science* **256**, 1555–1560 (1992).
- Savina, A., Furlán, M., Vidal, M. & Colombo, M. I. Exosome release is regulated by a calcium-dependent mechanism in K562 cells. *J. Biol. Chem.* **278**, 20083–20090 (2003).
- Soo, C. Y. *et al.* Nanoparticle tracking analysis monitors microvesicle and exosome secretion from immune cells. *Immunology* **136**, 192–197 (2012).



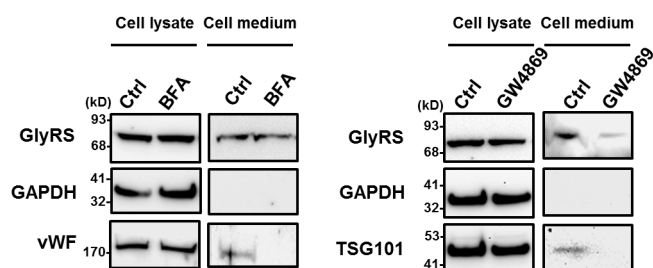
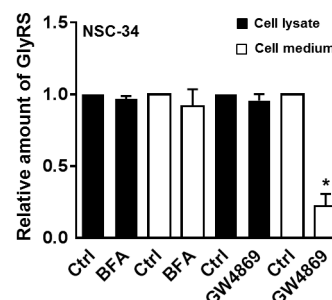
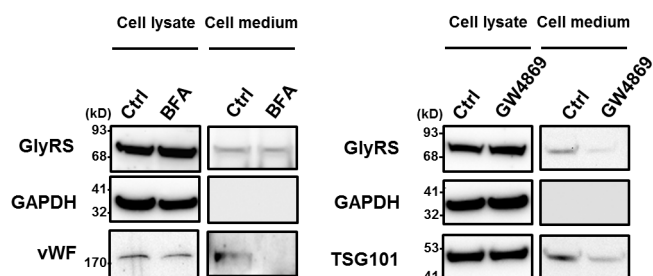
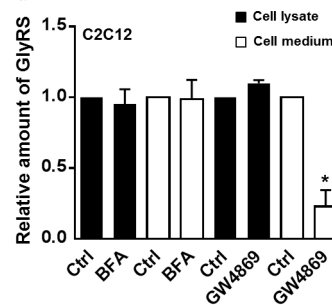
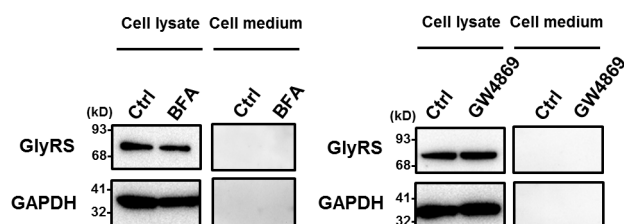
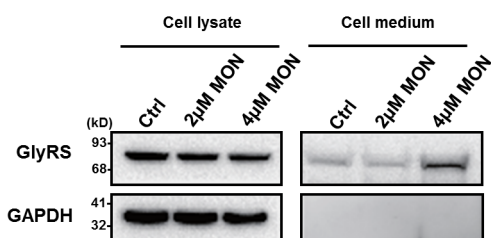
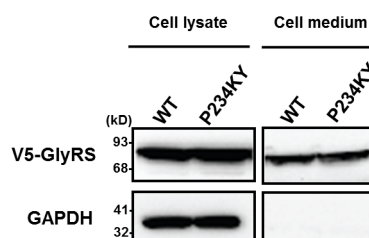
Extended Data Figure 1 | Hydrogen–deuterium exchange analysis to compare GlyRS^{CMT2D} (P234KY) and GlyRS^{WT} in solution. A global increase (15%) in deuterium incorporation for the mutant GlyRS was observed,

indicating overall structural opening. The regions having significant changes (>10%) in deuterium incorporation are highlighted under the human cytosolic GlyRS sequence with different colour codes (see box).



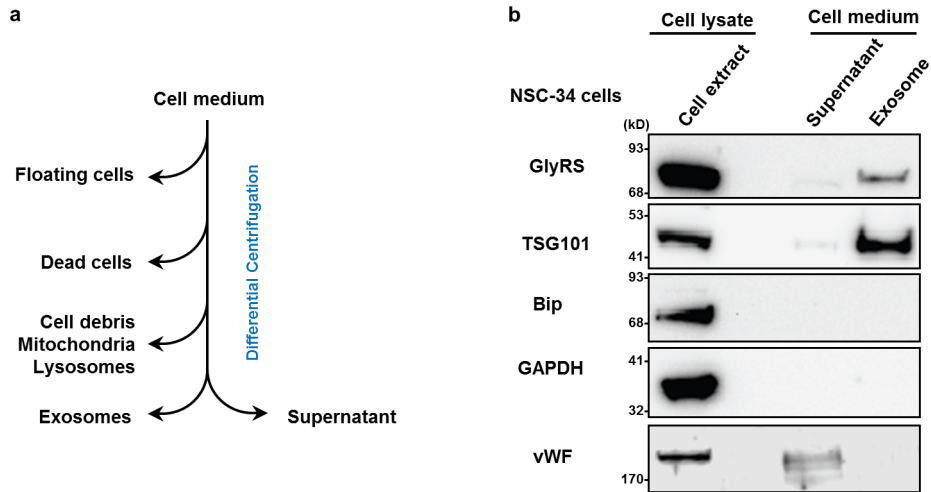
Extended Data Figure 2 | Characterization of the binding activity of GlyRS^{CMT2D}. **a**, *In vitro* pull-down of GlyRS^{CMT2D} (P234KY) proteins with the ectodomains of Nrp1, TrkB, Dcc, Robo1 and Unc5C proteins. Note the much stronger binding of GlyRS^{CMT2D} with Nrp1 compared with other receptors. GlyRS was detected by immunoblot with anti-GlyRS antibody; similar amounts of input receptors were visualized by Coomassie blue staining. **b**, *In vitro* pull-down of GlyRS^{CMT2D} proteins with the ectodomain of Nrp1. In addition to

L129P and P234KY, direct binding to Nrp1 was detected for E71G and G240R GlyRS^{CMT2D}. **c**, GST pull-down to confirm that b1 domain of Nrp1 is the main binding site of GlyRS^{CMT2D}. The amount of GST and GST fusion proteins used for GlyRS^{CMT2D} binding was visualized by Ponceau staining. **d, e**, *In vitro* pull-down assay showing the mutual competition between GlyRS^{CMT2D} (L129P) and VEGF-A₁₆₅ for Nrp1 binding.

a NSC-34 motor neurons**b****c C2C12 myotubes****d****e C2C12 myoblasts****f NSC-34 motor neurons****g Cos-7****Extended Data Figure 3 | Detection of GlyRS proteins in the cell medium.**

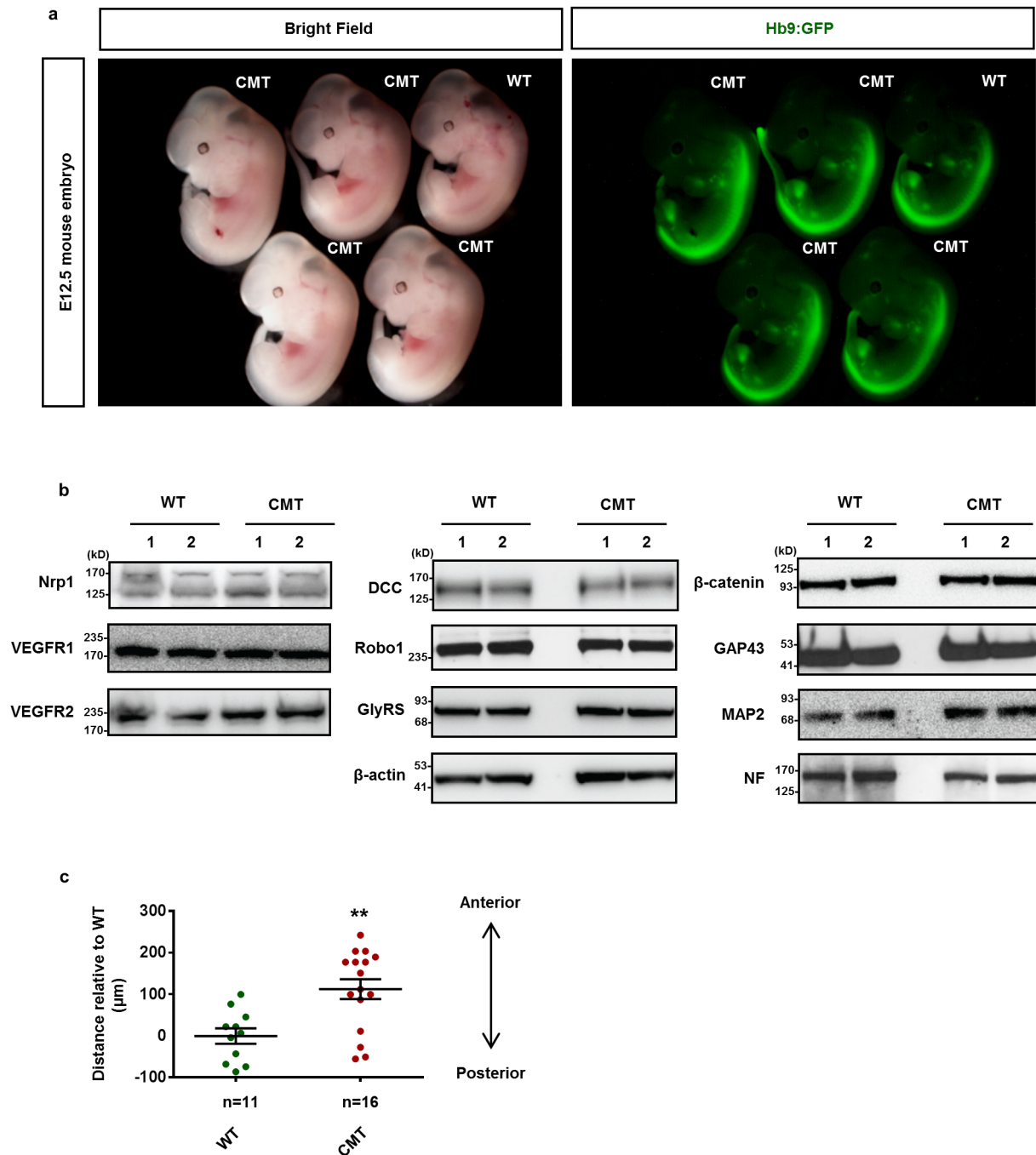
a, c, e, Western blot analysis of the GlyRS protein levels in NSC-34 motor neurons (**a**), C2C12 cell-differentiated myotubes (**c**) and undifferentiated C2C12 myoblasts (**e**). The observation that differentiated myotubes also secrete GlyRS raises the possibility that muscles, which are directly innervated by the peripheral motor neurons, might contribute to the disease pathology. The level of GlyRS proteins in cell medium is diminished by application of the exosome-pathway inhibitor GW4869, but not by brefeldin A (BFA), an inhibitor of the classical endoplasmic reticulum (ER)-to-Golgi secretory pathway. GAPDH (cytoplasmic protein), vWF (secretory protein through ER-Golgi pathway) and TSG101 (exosomal protein) are used as controls. **b, d,** Quantification of

GlyRS protein level indicated in **a, c**. Data are presented as the mean \pm s.e.m. of three independent experiments (* $P < 0.05$, t -test). **f,** Western blot analysis of the GlyRS protein level in NSC-34 motor neurons. The level of GlyRS proteins in the cell medium is increased by the treatment of monensin (MON), an activator for microvesicle release by regulating the intracellular calcium level^{40,41}. Vehicle-treated cells were used as control (Ctrl). **g,** Western blot analysis of the GlyRS protein level in Cos7 cells transfected with plasmids encoding GlyRS^{WT} and GlyRS^{CMT2D} (P234KY). The expression of GlyRS proteins was detected by immunoblot with antibody to V5 epitope tag. GAPDH was used as control. Note the similar level of GlyRS^{CMT2D} and GlyRS^{WT} in the media of transfected Cos7 cells.



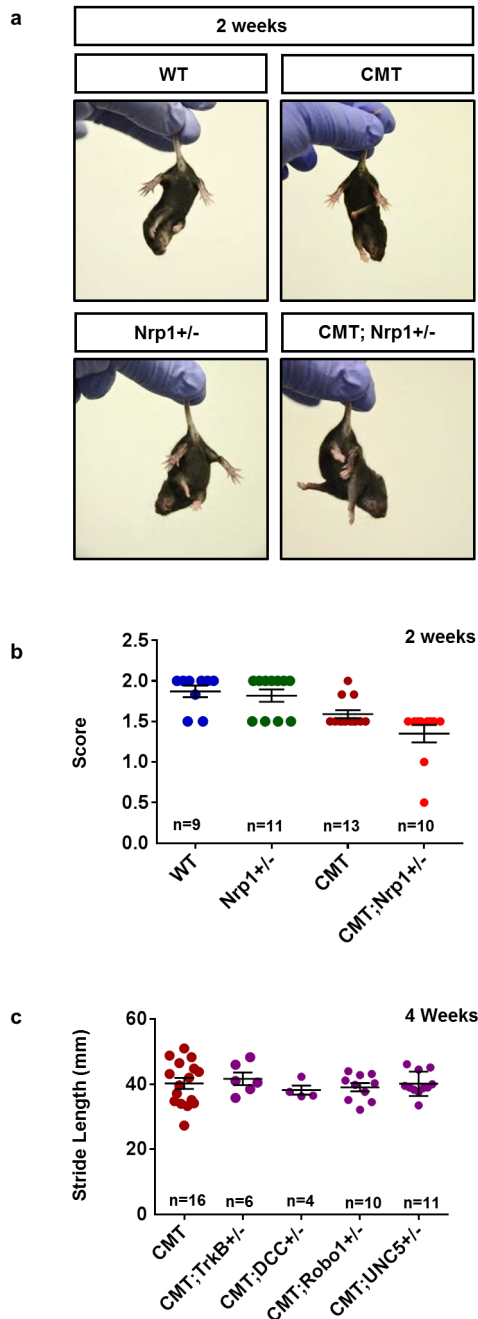
Extended Data Figure 4 | Detection of GlyRS proteins in exosome-enriched fractions. **a**, Diagram showing the procedure of exosome separation from the cell medium of NSC-34 cells by differential centrifugation. See Methods for details. **b**, Western blot analysis of proteins associated with various fractions. GlyRS proteins were detected in the exosome-enriched fractions but not in

supernatant fractions. The quality of the exosome preparation was controlled by detection of TSG101 (exosomal protein), Bip (ER-associated protein), GAPDH (cytoplasmic protein), and vWF (secretory protein through ER–Golgi pathway).

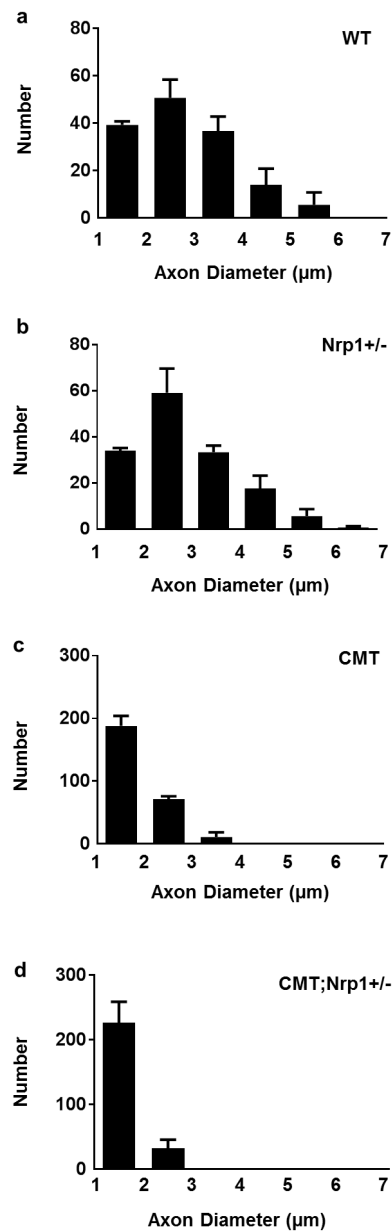


Extended Data Figure 5 | CMT2D mutant embryos have overall normal morphology but exhibit facial motor neuron migration defects. **a**, Lateral view of wild-type and CMT2D mutant embryos at E12.5. Motor neurons are specifically labelled by a transgenic fluorescence reporter, Hb9:GFP (green). Note overall normal morphology of CMT2D mutant embryos (CMT) compared with their littermate controls (wild type). **b**, Western blot analysis of protein expression in E12.5 mouse neural tissues. The expression levels of various neuronal proteins appear normal in CMT2D mutants compared with

their littermate controls. **c**, Considering CMT2D mutants show varying degrees of morphological change of facial motor nucleus, the facial motor neuron migration phenotype is quantified by measuring the relative distance of the facial motor nucleus between wild-type and CMT littermate embryos (each dot represents one facial motor nucleus, $n = 6$ embryos for wild type; $n = 8$ embryos for CMT2D). We find that the migration of facial motor neurons is significantly disrupted in CMT embryos. Data are presented as the mean \pm s.e.m. $**P < 0.01$ (t -test).

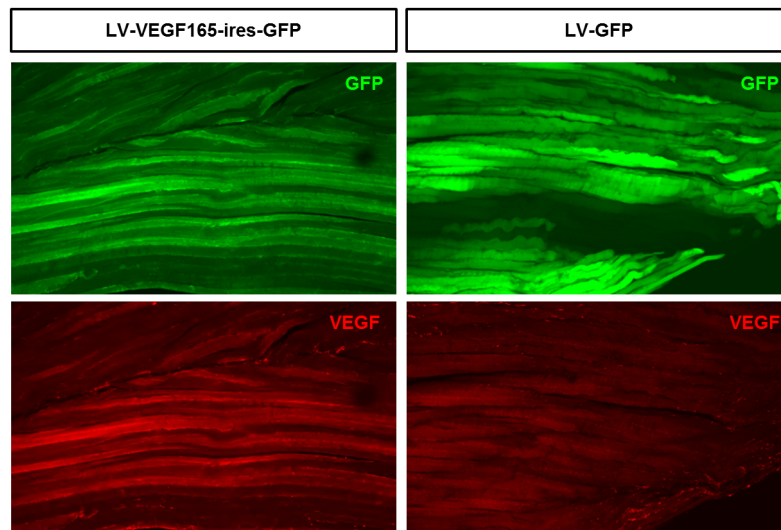


Extended Data Figure 6 | Genetic interaction between *Gars* and *Nrp1* in the early stage of CMT2D. **a, b,** Hindlimb extension test of wild-type and mutant animals at 2 weeks. Note that two out of nine *Gars*^{CMT2D};*Nrp1*^{+/-} (CMT;*Nrp1*^{+/-}) mutants exhibit hindlimb weakness with significantly lower scores compared with *Gars*^{CMT2D} (CMT), *Nrp1*^{+/-} and wild-type littermate controls. **c,** Comparison of stride lengths in different CMT2D mutant mice at 4 weeks old: *Gars*^{CMT2D} (CMT), *Gars*^{CMT2D};*TrkB*^{+/-} (CMT;*TrkB*^{+/-}), *Gars*^{CMT2D};*Dcc*^{+/-} (CMT;*Dcc*^{+/-}), *Gars*^{CMT2D};*Robo1*^{+/-} (CMT;*Robo1*^{+/-}), and *Gars*^{CMT2D};*Unc5C*^{+/-} (CMT;*Unc5C*^{+/-}). No significant differences were observed between compound heterozygotes and their littermate controls (CMT).



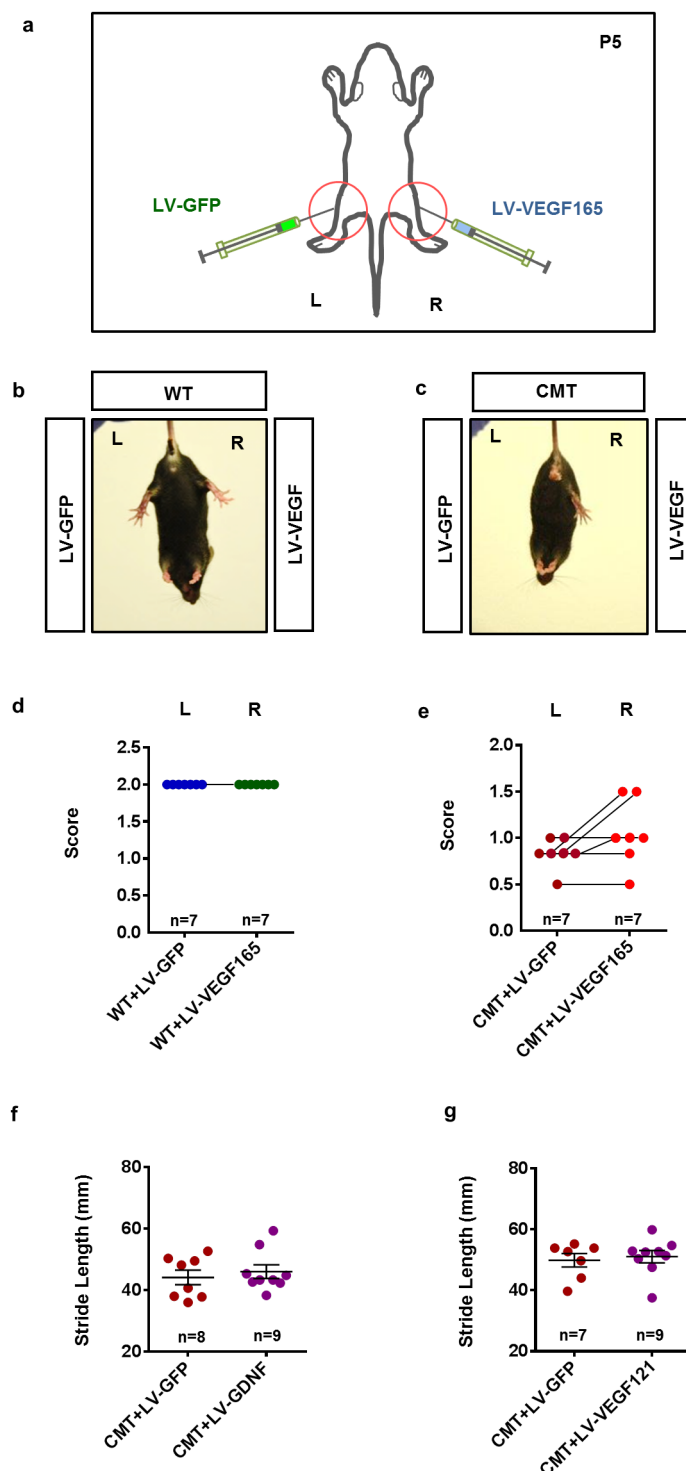
Extended Data Figure 7 | Axonal dystrophy in CMT2D mice.

a–d, Histograms showing the axonal diameter frequencies in the sciatic nerves of 4-week-old wild-type (**a**), *Nrp1*^{+/-} (**b**), *Gars*^{CMT2D} (CMT; **c**), and CMT;*Nrp1*^{+/-} (**d**) mutant mice. *n* = 3 mice per group. Note the decreased numbers of larger-diameter axons in CMT;*Nrp1*^{+/-} mutants compared with CMT, *Nrp1* heterozygous, and wild-type controls.



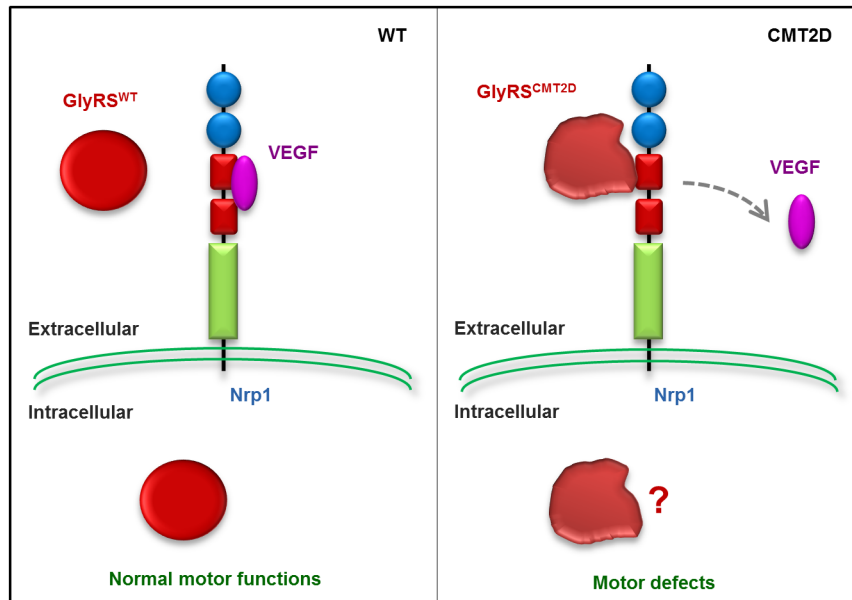
Extended Data Figure 8 | Expression level of VEGF in mouse muscles. The expression level of VEGF proteins in muscle fibres of mice injected with lentivirus expressing LV-VEGF₁₆₅-IRES-GFP versus LV-GFP was determined

by immunostaining with anti-VEGF antibodies. Note that the expression level of VEGF in LV-VEGF-infected muscles is significantly higher than in LV-GFP-infected control groups.



Extended Data Figure 9 | VEGF treatment retains limb strength in CMT2D mice. **a**, Diagram showing that lentiviral vectors encoding GFP (LV-GFP) or VEGF-A₁₆₅ (LV-VEGF₁₆₅) are injected unilaterally into each hindlimb of the same GlyRS^{CMT2D} mutant mouse at P5. **c**, **e**, At 5 weeks, LV-GFP-injected legs (L, left) of CMT2D animals have largely lost their ability to extend, while LV-VEGF165-treated legs (R, right) retained more limb strength with significantly higher scores in the hindlimb extension test (three out of seven animals). $P < 0.05$ (permutation test). **b**, **d**, No significant difference was

observed between both injected legs of wild-type animals in the hindlimb extension test. **f**, **g**, GDNF and VEGF-A₁₂₁ treatments fail to improve stride length in CMT2D mice. Walking strides of 2-month-old CMT2D mice bilaterally injected with lentiviral vectors (LV) encoding GFP, GDNF or VEGF-A₁₂₁. No significant difference of hindlimb stride length was observed between animals treated with LV-GDNF, LV-VEGF-A₁₂₁, and LV-GFP controls.



Extended Data Figure 10 | A simplified model for the neomorphic binding activity of GlyRS^{CMT2D}. Left, GlyRS^{WT} is a multifunctional protein with both intracellular and extracellular distributions. VEGF–Nrp1 signalling is an essential pathway for survival and function of motor neurons. (Note that VEGF may also act synergistically with other trophic factors, and/or maintain motor function indirectly by acting on Nrp1 receptors on non-motor neurons.)

Right, CMT2D mutations alter the conformation of GlyRS, enabling GlyRS^{CMT2D} to bind Nrp1. This aberrant interaction antagonizes the binding of VEGF to Nrp1, contributing to motor defects in CMT2D. Our results do not exclude the possibility that GlyRS^{CMT2D} may also interact with other extracellular and/or intracellular targets, related to CMT2D pathology.

Yap-dependent reprogramming of Lgr5⁺ stem cells drives intestinal regeneration and cancer

Alex Gregorieff¹, Yu Liu^{1,2}, Mohammad R. Inanlou¹, Yuliya Khomchuk¹ & Jeffrey L. Wrana^{1,2}

The gut epithelium has remarkable self-renewal capacity that under homeostatic conditions is driven by Wnt signalling in Lgr5⁺ intestinal stem cells (ISCs)¹. However, the mechanisms underlying ISC regeneration after injury remain poorly understood. The Hippo signalling pathway mediates tissue growth and is important for regeneration^{2,3}. Here we demonstrate in mice that Yap, a downstream transcriptional effector of Hippo, is critical for recovery of intestinal epithelium after exposure to ionizing radiation. Yap transiently reprograms Lgr5⁺ ISCs by suppressing Wnt signalling and excessive Paneth cell differentiation, while promoting cell survival and inducing a regenerative program that includes Egf pathway activation. Accordingly, growth of Yap-deficient organoids is rescued by the Egfr ligand epi-regulin, and we find that non-cell-autonomous production of stromal epi-regulin may compensate for Yap loss *in vivo*. Consistent with key roles for regenerative signalling in tumorigenesis, we further demonstrate that Yap inactivation abolishes adenomas in the *Apc*^{Min} mouse model of colon cancer, and that Yap-driven expansion of *Apc*^{-/-} organoids requires the Egfr module of the Yap regenerative program. Finally, we show that *in vivo* Yap is required for progression of early *Apc* mutant tumour-initiating cells, suppresses their differentiation into Paneth cells, and induces a regenerative program and Egfr signalling. Our studies reveal that upon tissue injury, Yap reprograms Lgr5⁺ ISCs by inhibiting the Wnt homeostatic program, while inducing a regenerative program that includes activation of Egfr signalling. Moreover, our findings reveal a key role for the Yap regenerative pathway in driving cancer initiation.

To examine Yap function in intestinal epithelial cells we first analysed Yap localization. This revealed cytoplasmic localization in most cells but absence in Paneth cells (Fig. 1a and Extended Data Fig. 1a). Upon exposure to whole-body irradiation (12 Gy), Yap was predominantly cytoplasmic with certain cells displaying weak nuclear accumulation at 1 day post-irradiation (dpi), was mostly nuclear at 2 dpi, and by 4 dpi returned to a prominent cytoplasmic localization (Fig. 1a). Analysis of irradiated *Yap*^{Δ/Δ} mice revealed strongly reduced crypt proliferation at 3 dpi (Fig. 1b and Extended Data Fig. 1b), contrasting with a previous study performed at 7 dpi that suggested Yap suppresses crypt regeneration⁴. However, we found that while 6 dpi *Yap*^{+/-} crypts had completed regeneration, surviving *Yap*^{Δ/Δ} or *Yap*^{Δ/Δ}; *Taz*^{Δ/Δ} crypts were still recovering and thus appeared enlarged and hyperproliferative (Extended Data Fig. 1b, c). In mosaic Yap-deficient intestines at 6 dpi (see Methods), surviving wild-type and Yap-mutant crypts were similarly enlarged and indistinguishable (Extended Data Fig. 1d). To explore Yap requirement further, we also deleted Yap in Lgr5⁺ ISCs, which are essential for regeneration⁵, using tamoxifen-inducible *Lgr5-creERT* in a *Rosa26-lacZ* background (*Yap*^{Δ/Δ}; *Lgr5-cre*)⁶. Twenty-four hours after induction, *Yap*^{Δ/Δ}; *Lgr5-cre* mice were irradiated (12 Gy) (Fig. 1c) and the fate of Yap-deficient Lgr5⁺ ISCs was monitored by Yap and β-gal double immunostaining (Fig. 1d). In untreated controls, 88% of β-gal-expressing crypts (517 of 587) lacked Yap, whereas in irradiated mice only 23% (91 of 394)

lacked Yap (Fig. 1e). Thus, most post-irradiation β-gal⁺ crypts arose from Lgr5⁺ clones that escaped Yap deletion. In agreement, *Yap*^{Δ/Δ} crypts displayed increased apoptosis at 1 dpi, while proliferation was unaffected (Extended Data Fig. 1e, f). Thus, Yap has an important early role in maintaining the ISC pool during intestinal regeneration, whereas the late hyperproliferative phase is independent of Yap and Taz.

To understand Yap function in ISCs, we used *ex vivo* organoids⁷, which revealed severe impairment in crypt formation upon Yap deletion (Extended Data Fig. 2a, b). Similar to post-irradiated crypts *in vivo*, Yap was cytoplasmic in differentiated zones (Extended Data Fig. 2c), but in crypts was both cytoplasmic and nuclear (Extended Data Fig. 2c). Because organoids model aspects of crypt regeneration, we assessed Yap transcriptional programs (Methods and Extended

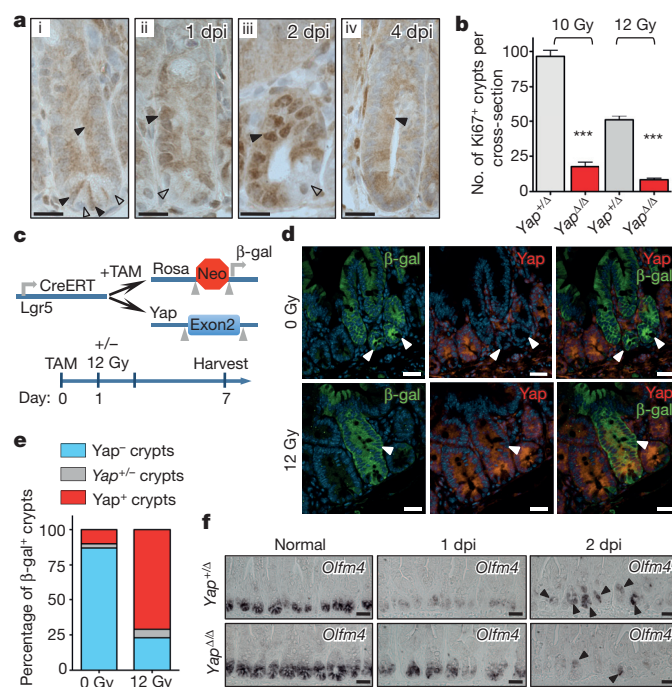


Figure 1 | Yap deficiency impairs crypt regeneration. **a**, Yap staining of intestines from untreated (i) and 12 Gy irradiated mice at indicated times (ii–iv) shows crypt base columnar, transit amplifying (filled arrowheads) and Paneth cells (open arrowheads). **b**, Quantification of number of surviving Ki67⁺ crypts in *Yap*^{+/Δ} versus *Yap*^{Δ/Δ} mice at 3 dpi (10 Gy and 12 Gy). *n* > 12; *n* = scored sections (see Methods); error bars indicate s.e.m.; ****P* < 0.0001. **c**, Design for tracing Yap-deficient Lgr5⁺ ISCs. **d**, Yap and β-gal immunostainings from untreated and irradiated (12 Gy) *Yap*^{Δ/Δ} mice (7 dpi). **e**, Survival of β-gal⁺ and either Yap⁺, Yap⁻, or mosaic Yap^{+/-} crypts was quantified (see Methods). **f**, *In situ* hybridization (ISH) analysis of ISC marker *Olfm4* in *Yap*^{+/Δ} and *Yap*^{Δ/Δ} intestines (*n* = 3). Scale bars: **a**, 35 μm; **d**, **f**, 70 μm.

¹Centre for Systems Biology, Lunenfeld Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, Ontario M5G 1X5, Canada. ²Department of Molecular Genetics, University of Toronto, Ontario M5S 1A8, Canada.

Data Fig. 3a), including in a Yap overexpression model that does not alter homeostasis *in vivo* but impairs organoid morphogenesis (Extended Data Fig. 3a–c). This revealed that Yap repressed Wnt targets and ISC signatures (for example, *Lgr5*, *Olfm4*, *Nkd1*, *Axin2*, *Ephb3* and *Aqp4*; see Supplementary Table 1), consistent with Yap and Taz inhibiting Wnt signalling (Extended Data Fig. 3d–f)^{4,8–10}. Accordingly, the ISC marker *Olfm4* was downregulated at 1 dpi *in vivo*, as reported previously^{11–13}, but persisted in *Yap*^{Δ/Δ} crypts (Fig. 1f). Furthermore, while control crypts recovered *Olfm4*⁺ cells by 2 dpi, in *Yap*^{Δ/Δ} mice *Olfm4* expression sharply declined (Fig. 1f), indicating ISC loss. Thus, Yap paradoxically promotes long-term ISC maintenance by suppressing Wnt signalling during early crypt regeneration.

We next asked how early inhibition of Wnt and *Olfm4*⁺ ISCs by Yap ultimately drives regeneration. High Wnt activity sensitizes ISCs to DNA damage and p53-dependent apoptosis¹³, consistent with increased apoptosis observed in 1 dpi *Yap*^{Δ/Δ} crypts (Extended Data Fig. 1e). In addition, while Wnt promotes self-renewal of *Lgr5*⁺ ISCs, it also drives Paneth cell differentiation^{14,15}. Notably, Yap also repressed Paneth markers such as *Wnt3*, *Spdef*, *Ccl6* and *Kit* (Extended Data Fig. 3d–f), and Yap-deficient crypts displayed marked cryptdin 1⁺ Paneth cell expansion by 2 dpi, with only rare *Olfm4*⁺ cells evident (Fig. 2a). Furthermore, while controls were strongly proliferative, with ectopic Paneth cells evident by 3 dpi, *Yap*^{Δ/Δ} crypts were almost entirely comprised of Paneth cells, with few proliferating cells (Extended Data Fig. 4a). These results are consistent with Paneth defects in mice deficient for Hippo components *Mst1* and *Mst2* (ref. 16). To test directly whether elevated Wnt contributes to these phenotypes, we lowered Wnt signalling in *Yap*^{Δ/Δ} organoids by reducing the levels of the Wnt agonist Rspo1 (ref. 17). This rescued *de novo* crypt formation and normalized Paneth cell differentiation, despite reduced overall growth (Fig. 2b, c and Extended Data Fig. 4b–d). Collectively, these results indicate that in the absence of Yap, high Wnt signalling drives ectopic Paneth cell differentiation that together with elevated apoptosis conspires to exhaust the ISC pool and block crypt regeneration.

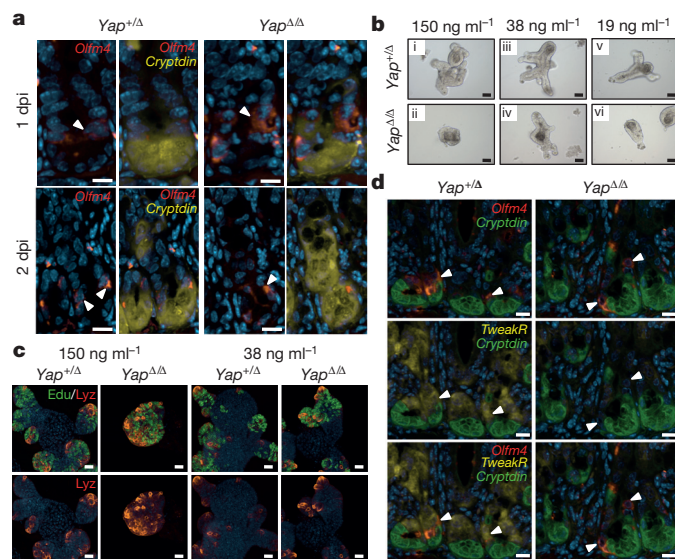


Figure 2 | Yap prevents excessive differentiation into Paneth cells during regeneration. **a**, Fluorescence ISH on irradiated *Yap*^{+/Δ} and *Yap*^{Δ/Δ} intestines using *Olfm4* and Paneth cell marker cryptdin 1 ($n = 3$). **b**, *Yap*^{+/Δ} and *Yap*^{Δ/Δ} organoids cultured in reducing Rspo1 concentrations (as indicated) for 3 days ($n = 4$ independent cultures). **c**, Lysozyme stainings and Edu incorporation in representative organoids cultured at 150 and 38 ng ml⁻¹ Rspo1. **d**, Fluorescence ISH to detect *TweakR*, *Olfm4* and cryptdin 1 expression in irradiated (2 dpi; 12 Gy) *Yap*^{+/Δ} and *Yap*^{Δ/Δ} intestines ($n = 3$). Scale bars: **a**, **c**, **d**, 35 μm; **b**, 70 μm.

Nuclear accumulation of Yap further suggested that a Yap-driven program contributes to regeneration. We identified numerous Yap-induced genes such as *TweakR* (also called *Tnfrsf12a*), *Areg*, *Edn1*, *Clu* and *Il1rn* (Supplementary Table 1 and Extended Data Fig. 3d–f) that are implicated in tissue repair, inflammation and cancer^{18,19}. In post-irradiated crypts, this program was strongly induced in a Yap-dependent manner, as well as in mice lacking *Lats1* and *Lats2*, which drives nuclear Yap (Extended Data Figs 5 and 6a). Furthermore, we observed Yap-dependent co-expression of *TweakR* and *Edn1* in regenerating *Olfm4*⁺ ISCs (Fig. 2d and Extended Data Fig. 6b), suggesting that Yap reprograms ISCs by transiently suppressing Wnt signalling, while concomitantly promoting the regenerative program. Egfr ligands (*Areg*, *Hbgef* and *Ereg*) were prominent in the Yap regenerative program and were of interest, because Egfr signalling is linked to ISC maintenance and Yap function^{20–25}. Interestingly, treatment of *Yap*^{Δ/Δ} organoids with Egfr ligands showed that exogenous Ereg rescued morphogenesis and suppressed apoptosis (Fig. 3a, b and Extended Data Fig. 7a), whereas additional Egf, a component of organoid growth medium, or Areg did not. The reasons for ligand specificity are unclear but may reflect non-redundant activities *in vivo*²⁶. Notably, stromal Ereg protects against intestinal damage and drives colitis-associated cancer^{27,28}. After irradiation, some *Yap*^{Δ/Δ} crypts survive and repopulate the gut, suggesting that stromal Ereg might rescue a handful of mutant crypts. Indeed, while *Ereg* was detected in certain cells of regenerating epithelia, it was also abundant in the underlying stroma (Extended Data Fig. 7b, c). Importantly, in *Yap*^{Δ/Δ} mice, stromal *Ereg* expression was even higher, while expression of other stromal growth factors was unchanged (Extended Data Fig. 7b, d). Thus, elevated stromal Ereg may compensate for Yap loss in some crypts that eventually regenerate the gut. Since reducing Wnt signalling also rescues *Yap*^{Δ/Δ} crypts *in vitro*, our results suggest that *in vivo* Yap coordinates Wnt antagonism and Egfr signalling to synergistically mediate efficient crypt regeneration.

To explore whether Yap regenerative responses contribute to cancer, we examined intestinal tumorigenesis models driven by loss of the tumour suppressor *Apc*. In *Apc*^{Min} adenomas and *Apc*^{Δ/Δ} organoids, Yap displayed nuclear localization (Extended Data Fig. 8a and Fig. 3c) and Yap deletion virtually abolished polyp formation in *Apc*^{Min} mice, markedly extending lifespans (Fig. 3d). Accordingly, Yap loss also blocked growth of *Apc*^{Δ/Δ} organoids and induced apoptosis (Fig. 3c, panel ii), while overexpression stimulated growth (Fig. 3e, panels i and ii). Thus, deregulated Wnt signalling caused by spontaneous *Apc* loss synergizes with Yap to drive tumorigenesis. We next examined the effect of crypt hyperplasia on acute *Apc* loss driven by *Vil*-CreERT. In contrast to a recent report¹⁰, deletion of Yap, Taz, or both Yap and Taz, had no effect on crypt hyperplasia, and Yap remained primarily cytoplasmic, indicating that *Apc* deletion *in vivo* is insufficient to drive nuclear localization of Yap (Extended Data Fig. 8b–d). In summary, Yap is required for adenoma formation but is dispensable for the proliferative burst associated with acute *Apc* loss.

Our results suggest that *Apc* deletion short circuits Yap inhibition of Wnt signalling, thus synergizing with Yap to drive polyp formation. Indeed, *Apc*^{Δ/Δ} organoids displayed strong induction of Wnt targets that were only weakly affected by Yap loss, consistent with Yap antagonism upstream of *Apc* (Extended Data Fig. 9a)⁸. However, Yap-induced genes that included Egfr agonists, and Yap suppression of Paneth markers, were unaffected by *Apc* loss, indicating that Yap also suppresses Paneth cells independent of its anti-Wnt effects (Extended Data Fig. 9a). Of note, *Apc*^{Δ/Δ} organoids do not require exogenous Egfr (Fig. 3c), so we explored whether this is due to the Yap-induced Egfr module. Egfr (PD153053) or Mek (U0126) inhibition abrogated growth of *Apc*^{Δ/Δ} spheroids and Yap-transgene-driven *Apc*^{Δ/Δ} spheroids (Fig. 3c, panel iv, and 3e, panels iii–vi). Furthermore, Yap loss diminished, whereas upregulated Yap enhanced, phospho-Erk1/2 levels (Extended Data Fig. 9b), and ectopic Ereg rescued *Yap*^{Δ/Δ}; *Apc*^{Δ/Δ} spheroid growth (Extended Data Fig. 9c, d). The Egfr module of

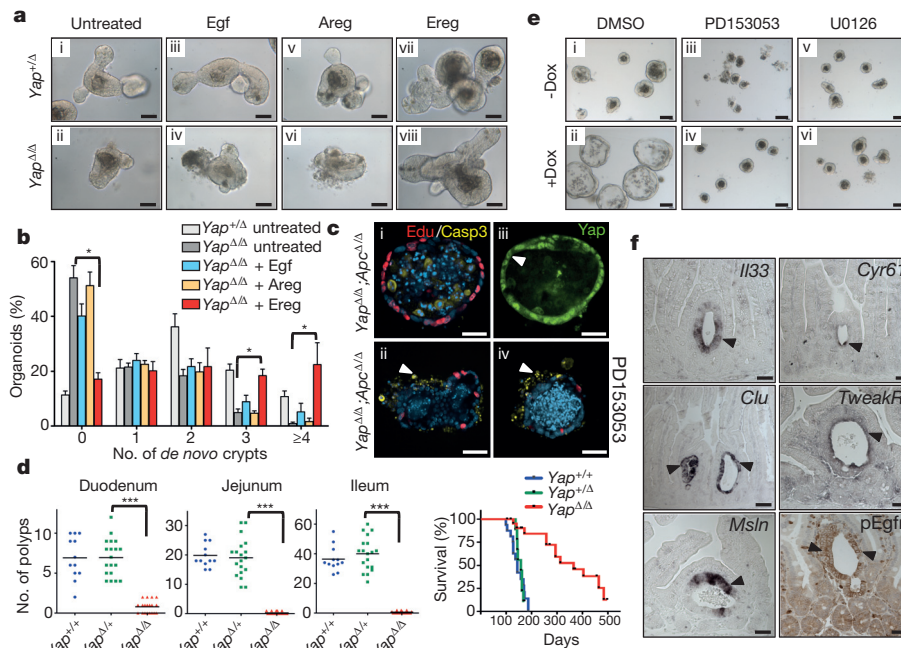


Figure 3 | Yap/Egfr signalling in intestinal regeneration and tumorigenesis. **a**, Yap^{+/Δ} and Yap^{Δ/Δ} organoids were supplemented with 0.5 μg ml⁻¹ of recombinant Egfr, Areg, or Ereg. **b**, Percentage of organoids with 0, 1, 2, 3, or ≥4 *de novo* crypts formed after 4 days of culture (*n* = 4 independent cultures, **P* < 0.05). Error bars indicate s.e.m. **c**, Yap^{+/Δ};Apc^{Δ/Δ} and Yap^{Δ/Δ};Apc^{Δ/Δ} organoids grown for 5 days (i–ii). Proliferation and apoptosis were examined by EdU incorporation and active caspase 3, respectively. Yap expression is shown in panel iii. Treatment of Apc^{Δ/Δ} organoids with Egfr inhibitor (PD153035) mimics loss of Yap (iv) (*n* = 5). **d**, Tumour load in age-matched Apc^{Min}, Yap^{+/+} (*n* = 12), Yap^{+/Δ} (*n* = 19) and Yap^{Δ/Δ} (*n* = 19) mice (*n* = number of mice per group; error bars indicate s.e.m.; ****P* < 0.0001). The right panel shows survival curves of Yap^{+/+} (*n* = 9), Yap^{+/Δ} (*n* = 16) and Yap^{Δ/Δ} (*n* = 23) mice. **e**, YapTg;Apc^{Δ/Δ} organoids were cultured 4 days with or without doxycycline and treated with inhibitors: PD153035 (Egfr) or U0126 (Mek) (*n* = 5). **f**, Expression of Il33, Cyr61, Msln, TweakR and phospho-Egfr (Tyr-1092) in Apc mutant aberrant crypt foci. Scale bars, 70 μm.

Yap regenerative signalling is thus an important target of the pro-tumorigenic effects of Yap.

To examine how Yap-driven regenerative signalling contributes to Apc mutant adenoma formation, we generated *Yap*;Apc^{ALgr5-cre} (*Yap*^{fl/fl};Apc^{fl/fl};Lgr5-creERT) mice, which allow timed induction of sporadic Yap-null and Apc-null ISCs that are tracked by staining Yap and the Wnt target gene Lef, respectively (Extended Data Fig. 10a, b and Methods). This yielded both Yap wild-type and null Apc mutant, Lef⁺ foci that were readily detected early, consistent with Yap and Taz being dispensable for expansion of acute Apc^{-/-} lesions (Extended Data Fig. 10c, panels i and ii). Furthermore, Yap was cytoplasmic in most, but not all, cells (Extended Data Fig. 10c, panel iii), further supporting that Apc loss is not sufficient to drive nuclear localization of Yap. However, at later stages, when Lef⁺ lesions matured to morphologically distinct aberrant crypt foci and adenomas, Yap was predominantly nuclear (Extended Data Fig. 10c, panel vi), which coincided with expression of the Yap regenerative signature, including Egfr pathway components (Fig. 3f). The rarity of adenomas in *Yap*;Apc^{ALgr5-cre} and *Yap*^{Δ/Δ};Apc^{Min} mice (data not shown) raised the question of why Yap mutant Lef⁺ foci undergo early demise. Apoptosis was not altered in Yap-deficient, Lef⁺ foci (Extended Data Fig. 10d), but Paneth cells were strongly enriched (Fig. 4a, b, panels i–iii, and Extended Data Fig. 10e, panels i–iii) with many lysozyme⁺, Yap mutant foci displaying weak Lef expression (Fig. 4b, panels iv–vi, and Extended Data Fig. 10e, panels iv–vi), indicating depletion into post-mitotic Paneth cells. Notably, mTOR inhibition also increases Paneth differentiation and reduces tumorigenesis in Apc models²⁹. We also examined Egfr activity in these early lesions. Although weaker than in mature lesions, where the Yap regenerative program is robustly expressed, Egfr activation was readily detected in Yap wild-type lesions, but significantly reduced in Yap mutants (Fig. 4c and Extended Data Fig. 10f). Thus, Yap regenerative signalling probably promotes progression of Apc^{-/-} foci to adenomas by suppressing Paneth differentiation and activating Egfr signalling.

These studies define a central role for the Yap effector of the Hippo pathway in reprogramming ISCs during intestinal regeneration and driving tumorigenesis (Fig. 4d). Yap protects the ISC pool by transiently suppressing Wnt signalling and the Lgr5⁺ ISC population, while blocking differentiation into Paneth cells and driving a pro-regenerative program critically dependent on induction of Egfr signalling. We showed that aberrant crypt morphogenesis in Yap-deficient organoids was overcome by either reducing Wnt signalling or by

addition of the Egfr ligand epiregulin. These two Yap-dependent processes may thus function in parallel and act synergistically *in vivo* to promote regeneration. Finally, we describe a mechanistic link

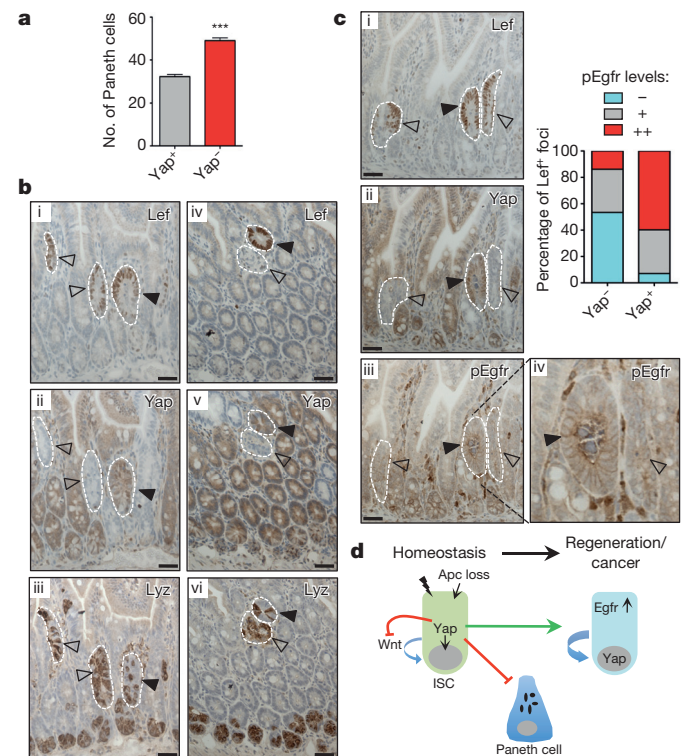


Figure 4 | Yap-dependent Egfr activation and suppression of Paneth cell differentiation in tumour initiating cells. **a**, Relative abundance of Paneth cells in Yap-positive (*n* = 207) and Yap-negative (*n* = 201) Lef⁺ foci from tamoxifen-induced *Yap*;Apc^{ALgr5-cre} mice (*n* = number of Lef⁺ foci; error bars indicate s.e.m.; ****P* < 0.0001, see Methods). **b**, Two sets of consecutive sections (i–iii and iv–vi) showing Lef, Yap and lysozyme (Lyz) staining of representative Lef⁺ foci. **c**, Lef, Yap and phospho-Egfr (i–iv) staining of consecutive sections from *Yap*;Apc^{ALgr5-cre} mice. Graph shows quantification of expression levels of phospho-Egfr in Yap positive and negative Lef⁺ foci (see Methods). Closed and open arrowheads indicate Yap positive and negative Lef⁺ foci, respectively. **d**, A model of Yap function in intestinal regeneration and cancer. Scale bars, 70 μm.

between tissue regeneration and cancer in which mutations in APC or β -catenin that short-circuit Yap suppression of Wnt⁸ synergize with Yap regenerative signalling and Paneth cell suppression to create early tumorigenic lesions.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 14 May 2014; accepted 7 August 2015.

Published online 21 October 2015.

- Clevers, H. The intestinal crypt, a prototype stem cell compartment. *Cell* **154**, 274–284 (2013).
- Johnson, R. & Halder, G. The two faces of Hippo: targeting the Hippo pathway for regenerative medicine and cancer treatment. *Nature Rev. Drug Discov.* **13**, 63–79 (2014).
- Yu, F. X., Meng, Z., Plouffe, S. W. & Guan, K. L. Hippo pathway regulation of gastrointestinal tissues. *Annu. Rev. Physiol.* **77**, 201–227 (2015).
- Barry, E. R. *et al.* Restriction of intestinal stem cell expansion and the regenerative response by YAP. *Nature* **493**, 106–110 (2013).
- Metcalfe, C., Klijavin, N. M., Ybarra, R. & de Sauvage, F. J. Lgr5⁺ stem cells are indispensable for radiation-induced intestinal regeneration. *Cell Stem Cell* **14**, 149–159 (2014).
- Barker, N. *et al.* Identification of stem cells in small intestine and colon by marker gene Lgr5. *Nature* **449**, 1003–1007 (2007).
- Sato, T. *et al.* Single Lgr5 stem cells build crypt-villus structures *in vitro* without a mesenchymal niche. *Nature* **459**, 262–265 (2009).
- Varelas, X. *et al.* The Hippo pathway regulates Wnt/ β -catenin signaling. *Dev. Cell* **18**, 579–591 (2010).
- Imajo, M., Miyatake, K., Iimura, A., Miyamoto, A. & Nishida, E. A molecular mechanism that links Hippo signalling to the inhibition of Wnt/ β -catenin signalling. *EMBO J.* **31**, 1109–1122 (2012).
- Azzolin, L. *et al.* YAP/TAZ incorporation in the β -catenin destruction complex orchestrates the Wnt response. *Cell* **158**, 157–170 (2014).
- Yan, K. S. *et al.* The intestinal stem cell markers Bmi1 and Lgr5 identify two functionally distinct populations. *Proc. Natl Acad. Sci. USA* **109**, 466–471 (2012).
- van Es, J. H. *et al.* Dll1⁺ secretory progenitor cells revert to stem cells upon crypt damage. *Nature Cell Biol.* **14**, 1099–1104 (2012).
- Tao, S. *et al.* Wnt activity and basal niche position sensitize intestinal stem and progenitor cells to DNA damage. *EMBO J.*, (2015).
- van Es, J. H. *et al.* Wnt signalling induces maturation of Paneth cells in intestinal crypts. *Nature Cell Biol.* **7**, 381–386 (2005).
- Andreu, P. *et al.* Crypt-restricted proliferation and commitment to the Paneth cell lineage following Apc loss in the mouse intestine. *Development* **132**, 1443–1451 (2005).
- Zhou, D. *et al.* Mst1 and Mst2 protein kinases restrain intestinal stem cell proliferation and colonic tumorigenesis by inhibition of Yes-associated protein (Yap) overabundance. *Proc. Natl Acad. Sci. USA* **108**, E1312–E1320 (2011).
- Farin, H. F., Van Es, J. H. & Clevers, H. Redundant sources of Wnt regulate intestinal stem cells and promote formation of Paneth cells. *Gastroenterology* **143**, 1518–1529 (2012).
- Karaca, G. *et al.* TWEAK/Fn14 signaling is required for liver regeneration after partial hepatectomy in mice. *PLoS ONE* **9**, e83987 (2014).
- Shao, J. & Sheng, H. Amphiregulin promotes intestinal epithelial regeneration: roles of intestinal subepithelial myofibroblasts. *Endocrinology* **151**, 3728–3737 (2010).
- Ren, F. *et al.* Hippo signaling regulates *Drosophila* intestine stem cell proliferation through multiple pathways. *Proc. Natl Acad. Sci. USA* **107**, 21064–21069 (2010).
- Powell, A. E. *et al.* The pan-ErbB negative regulator Lrig1 is an intestinal stem cell marker that functions as a tumor suppressor. *Cell* **149**, 146–158 (2012).
- Staley, B. K. & Irvine, K. D. Warts and Yorkie mediate intestinal regeneration by influencing stem cell proliferation. *Curr. Biol.* **20**, 1580–1587 (2010).
- Wong, V. W. *et al.* Lrig1 controls intestinal stem-cell homeostasis by negative regulation of ErbB signalling. *Nature Cell Biol.* **14**, 401–408 (2012).
- Yang, N. *et al.* TAZ induces growth factor-independent proliferation through activation of EGFR ligand amphiregulin. *Cell Cycle* **11**, 2922–2930 (2012).
- Zhang, J. *et al.* YAP-dependent induction of amphiregulin identifies a non-cell-autonomous component of the Hippo pathway. *Nature Cell Biol.* **11**, 1444–1450 (2009).
- Pastore, S., Mascia, F., Mariani, V. & Girolomoni, G. The epidermal growth factor receptor system in skin repair and inflammation. *J. Invest. Dermatol.* **128**, 1365–1374 (2008).
- Lee, D. *et al.* Epiregulin is not essential for development of intestinal tumors but is required for protection from intestinal damage. *Mol. Cell. Biol.* **24**, 8907–8916 (2004).
- Neufert, C. *et al.* Tumor fibroblast-derived epiregulin promotes growth of colitis-associated neoplasms through ERK. *J. Clin. Invest.* **123**, 1428–1443 (2013).
- Faller, W. J. *et al.* mTORC1-mediated translational elongation limits intestinal tumour initiation and growth. *Nature* **517**, 497–500 (2015).

Supplementary Information is available in the online version of the paper.

Acknowledgements We would like to thank K. Chan for performing RNA-seq analysis, M. Moran for advice on Egr analyses, and R. Bremner and L. Attisano for critical review of the manuscript. This work was supported by the CIHR (MOP-12860 and MOP-106672), the Terry Fox Research Institute, and the Krembil Foundation. J.L.W. is the Mary Janigan Chair in Experimental Therapeutics and the CIBC Chair in Breast Cancer Research.

Author Contributions Experiments were conceived and designed by A.G. and J.L.W. Experiments were performed by A.G., Y.L., M.R.I. and Y.K. pEgr staining was quantified by A.G. and J.L.W. Bioinformatic analysis of RNA-seq data was performed by Y.L. The manuscript was written by A.G. and J.L.W.

Author Information Data from RNA sequencing analysis have been deposited in the GEO repository under accession GSE66567. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.L.W. (wrana@lunenfeld.ca).

METHODS

No statistical methods were used to predetermine sample size.

Mice. *Yap^{Δ/Δ}* and *Taz^{Δ/Δ}* mice were generated by crossing *Yap* or *Taz* floxed mice³⁰ with the *villin-cre* line (Jackson Laboratory), the *villin-creERT2* line (S. Robine, Institut Curie-CNRS) or the *Lgr5-creERT* line (Jackson Laboratory). The *Rosa26-lox-STOP-lox-rtta-IRES-EGFP* and *Rosa26 lacZ* mouse lines were obtained from Jackson Laboratory. The *YapTg* transgenic line described in this study was generated by introducing a HA-tagged wild-type *Yap* cDNA downstream of 7 Tet-repressor elements in the pTRE2 vector (J. Whitsett, Cincinnati Children's Hospital Medical Center). The transgenic *Yap* construct was linearized and microinjected in ICR embryos. As shown in Extended Data Fig. 3a, activation of Cre deletes a *neo* cassette and allows for expression of the *rtTA* gene. In the presence of doxycycline the *rtTA* activates transcription of *HA-Yap*. *Apc* floxed mice were obtained from O. Sansom (Beatson Institute). *Lats1* and *Lats2* floxed alleles were obtained from R. Johnson (MD Anderson Cancer Center) and crossed with *villin-creERT2* mice to obtain *Lats1^{Δ/Δ}*; *Lats2^{Δ/Δ}* mice. To measure polyp formation, *Yap^{Δ/Δ}* mice were backcrossed to a Bl/6 background for 4 generations before crossing to *Apc^{Min/+}* mice. Polyps from *Yap^{+/+}* (*Yap^{+/+}*; *villin-cre*; *Apc^{Min/+}*), *Yap^{Δ/Δ}* (*Yap^{Δ/Δ}*; *villin-cre*; *Apc^{Min/+}*) and *Yap^{Δ/Δ}* (*Yap^{Δ/Δ}*; *villin-cre*; *Apc^{Min/+}*) mice were counted 16 weeks after birth or when animals appeared moribund. Survival of *Apc^{Min}* mice was measured by the number of days before mice were euthanized due to poor health. *In vivo* assays comparing control and *Yap* mutant animals were performed between age- and sex-matched pairs. No method of randomization was followed and no animals were excluded in this study. The investigators were not blinded to allocation during experiments and outcome assessment. Inducible Cre-mediated deletion of genes was performed by intraperitoneal injections of >5-week-old mice with 200 μ l tamoxifen in corn oil at 10 mg ml⁻¹. To create mosaic expression of *Yap*, *Yap^{fl/fl}*; *villin-creERT2* mice were induced with a single injection of 200 μ l of tamoxifen at a suboptimal dose typically between 0.5 and 2.0 mg ml⁻¹. For *in vivo* regeneration assays, mice were given a single dose of 10 or 12 Gy using a GammaCell 40 irradiator. Animals were maintained and handled under procedures approved by the Canadian Council on Animal Care.

Immunohistochemistry and *in situ* hybridization. The immunohistochemistry stainings and standard colorimetric *in situ* hybridization were carried out according to methods described elsewhere³¹. Staining experiments were repeated on independent tissue sections prepared from separate mice as indicated by *n* values in figure legends. The following primary antibodies were used for immunostaining: rat anti-Ki67 (Dako, Cat. no. M7249, 1:1,000), rabbit anti-*Yap*/*Taz* (Cell Signaling, Cat. no. 8418, 1:100), rabbit anti-*Yap* (Cell Signaling, Cat. no. 14074, 1:300), mouse anti-*Yap* (Santa Cruz, Cat. no. sc-101199, 1:100), rabbit anti-*Lef* (Cell Signaling, Cat. no. 2230, 1:300), phospho-Egfr (Tyr1092) (Abcam, Cat. no. ab40815, 1:300), anti-cleaved caspase-3 (Cell Signaling, Cat. no. 9664, 1:300) and anti-lysozyme (Dako, Cat. no. A0099, 1:1,000). Detection of primary antibodies was achieved using the Dako Envision plus system. Multi-colour fluorescence *in situ* hybridization with tyramide signal amplification (TSA) was done essentially as described elsewhere^{32–34}. In brief, RNA probes from hybridized sections were detected using appropriate hapten-specific HRP-conjugated antibodies (anti-digoxigenin-HRP (Roche, Cat. no. 11207733910, 1:500), anti-dinitrophenyl-HRP (PerkinElmer, Cat. no. NEL747A001KT, 1:300), and anti-fluorescein-HRP (Life Technologies, A21253, 1:500)). After overnight incubation with antibodies at 4°C (or 2 h at room temperature for anti-fluorescein-HRP detection of cryptin1) sections were washed in PBS, and rinsed twice in 100 mM borate pH 8.5 plus 0.1% BSA. TSA reaction was performed by applying 300 μ l per slide of the following mixture: 100 mM borate pH 8.5, 2% dextran sulfate, 0.1% Tween-20 and 0.003% H₂O₂, 450 μ g ml⁻¹ 4-iodophenol: 1:250 Tyramide product (that is, DyLight633-tyramide, Dylight488-tyramide, Dylight 555-tyramide). The TSA reaction was allowed to proceed for 20 min and then terminated by washing slides in 100 mM glycine pH 2.0 for 15 min. Sections were washed further in PBS for the next round of detection. To synthesize tyramide products, the following succinimidyl esters were used for conjugation with tyramine: DyLight 633 NHS-Ester (Thermo Scientific Cat#46414), DyLight 550 NHS-Ester (Thermo Scientific Cat#62262), DyLight 488 NHS-Ester (Thermo Scientific Cat. no. 46402). The synthesis reaction was carried out as described previously³².

The following *in situ* hybridization probes were obtained from the collection of MGC clones at the Lunenfeld Tanenbaum Research Institute: *TweakR* (BC025860), *Ly6cl* (BC092082), *Edn1* (BC029547), *Areg* (BC009138), *Ereg* (BC027838), *Il1rn* (BC042532), *Il33* (BC003847), *Msln* (BC023753) and *Cyr61* (BC066019). The *Olfr4* and cryptdin1 probes were a gift from H. Clevers (Hubrecht Institute).

Immunofluorescence. Before fixing organoids, 10 μ M Edu was added to the culture media for 1 h. Then organoids were fixed in 10% buffered formalin for

30 min, permeabilized in 0.5% Triton for 20 min and blocked in 2% BSA. Incorporated Edu was detected using the ClickIt EDU Imaging kit (Invitrogen) according to the manufacturer's instructions. The primary antibodies used for immunostaining were mouse anti-*Yap* (Santa Cruz, Cat # sc-101199, 1:100), mouse anti-HA (Sigma-Aldrich, Cat. no. H9658, 1:1,000), and chicken anti- β -gal (Abcam, Cat. no. ab9361, 1:300). The secondary antibodies used in immunostaining were: CF555-donkey anti-mouse (Biotium, Cat. no. 20037, 1:400) and CF647 donkey anti-rabbit (Biotium, Cat. no. 20047, 1:400). Organoids were counterstained with 4',6-diamidino-2-phenylindole dihydrochloride (DAPI) (Sigma-Aldrich) before mounting onto slides for visualization. Images were acquired using a 20 \times /NA oil immersion objective lens (HCX PL APO, Leica), an EM-CCD camera (ImagEM, Hamamatsu) on an inverted microscope (DMIRE2, Leica) with a spinning disk confocal scanner (CSU10, Yokogawa) and Velocity.

Quantification of *de novo* crypt formation. *De novo* crypts were scored as any protrusions, typically containing Paneth cells, budding from the initial sphere formed after seeding isolated crypts. Crypts were counted from bright-field images using Image J. At least four independent cultures derived from four different mice per genotype were used for quantification.

***In vivo* quantification.** Survival of crypts in Fig. 1 was determined by Ki67 staining of cross-sections of proximal portions of the small intestine at 3 days post-irradiation (10 Gy or 12 Gy). Values in Fig. 1b represent average number of fully labelled Ki67⁺ crypts per intestinal circumference based on counts from at least two sections per mouse and assays were repeated in 6 independent mice per genotype for both 10 Gy and 12 Gy treatments. The percentage of surviving *Yap*-positive versus negative *Lgr5⁺* ISC in Fig. 1d was performed by counting 587 β -gal⁺ crypts from a total of 7 untreated *Yap^{ΔLgr5-cre}* mice and 394 β -gal⁺ crypts from a total of 9 irradiated *Yap^{ΔLgr5-cre}* mice. In *Yap*; *Apc^{ΔLgr5-cre}* mice tumour initiating cells were visualized by staining for the Wnt target gene, *Lef*. As shown in Extended Data Fig. 10b, *Lef* is undetected in wild-type crypts and highly upregulated in *Apc*-null cells and thus serves as a robust marker of *Apc* deletion³¹. The percentage of Paneth cells in *Lef⁺* foci (Fig. 4a) was assessed by preparing consecutive sections stained for *Lyz*, *Lef* and *Yap*, respectively. Lysozyme-positive Paneth cells from a total of 207 *Yap* wild-type and 201 *Yap* mutant *Lef⁺* foci were counted from 5 *Yap*; *Apc^{ΔLgr5-cre}* mice (10–16 days after tamoxifen injection) using Image J and the percentage of total cells within the boundaries of a given *Lef⁺* lesion was calculated. Relative activation of *Egfr* was quantified in consecutive sections from 5 *Yap*; *Apc^{ΔLgr5-cre}* mice stained for *Lef*, *Yap* and phospho-Egfr. For assessing Phospho-Egfr, staining intensity in *Lef⁺* foci was assessed in a blinded fashion. For this, consecutive sections stained for *Yap* were masked from the observer scoring phospho-Egfr staining intensity. *Lef⁺* foci were scored as '+' if phospho-Egfr expression was elevated compared to wild-type adjacent crypts at comparable levels within the crypt-villus axis (see Extended Data Fig. 10f, panels xi and xii). *Lef⁺* foci were scored as '++' if staining intensity was very strong even relative to the stem cell compartment in normal crypts and/or displayed prominent apical staining (see *Yap*-positive foci in Fig. 4c, panel iv, and Extended Data Fig. 10f, panels v and vi). *Lef⁺* foci were scored as '-' if staining intensity was undetected or unchanged relative to adjacent wild-type crypts (see *Yap* mutant foci in Fig. 4c and Extended Data Fig. 10f). In Extended Data Fig. 1, caspase 3 and BrdU positive crypt cells were counted from at least six sections per mouse in 4 independent mice per genotype and expressed as a percentage of total crypt cells.

Statistics. All data are presented as average values with s.e.m. Mann-Whitney (two-tailed) *U*-test was used to determine statistical significance. Calculations were performed using GraphPad Prism 5 software.

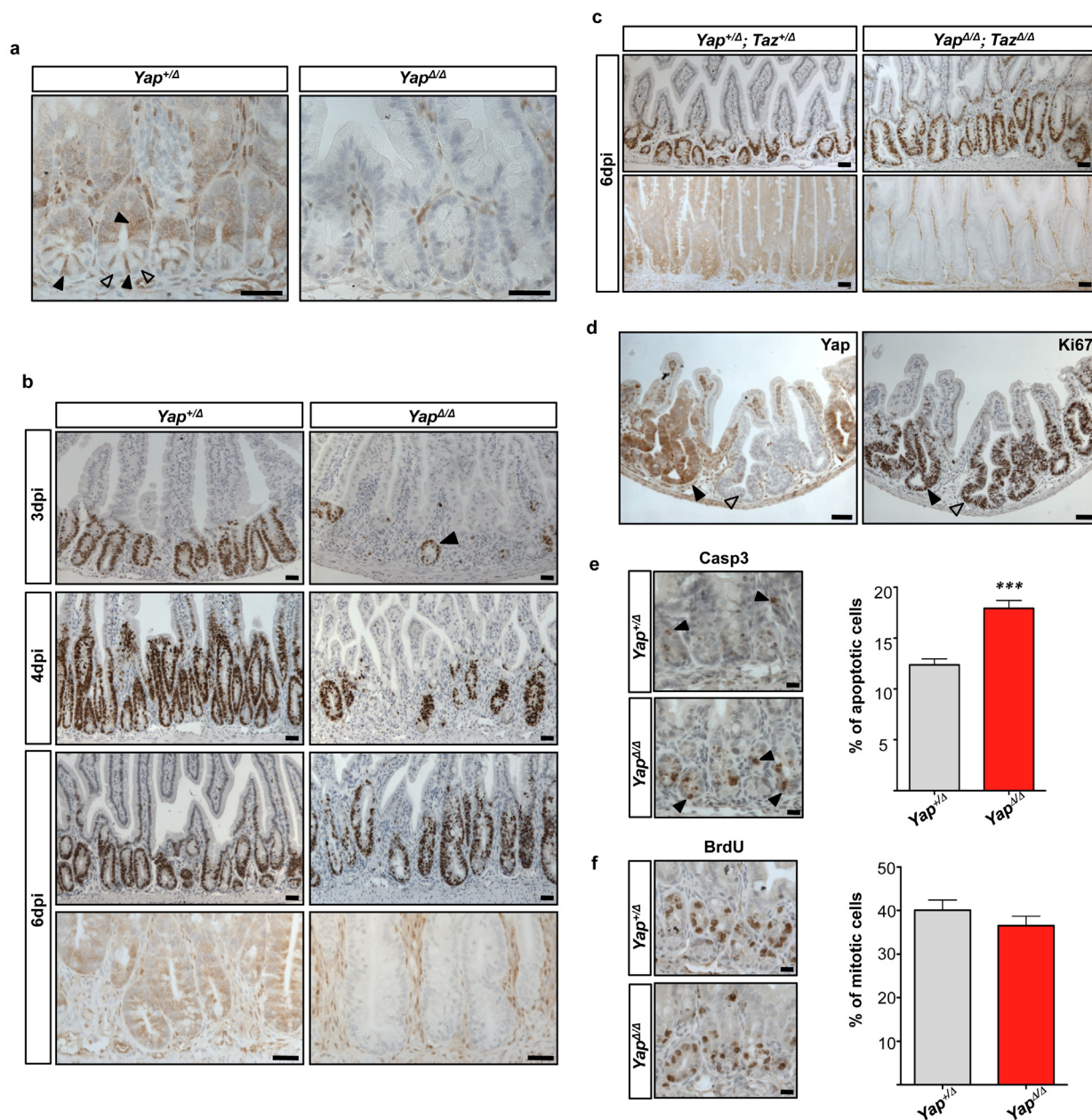
RNA-seq analysis. RNA was isolated from organoids cultured for 24 h after seeding in Matrigel. RNA samples were pooled from at least three organoid cultures derived from at least three independent mice per genotype (*Yap^{fl/fl}*; *villin-cre*, *Yap^{fl/fl}*; *villin-cre* and *YapTg*). Quality of RNA was verified by running samples on a Bioanalyzer. High-throughput sequencing was performed using the Illumina HiSeq 2000 at the Lunenfeld Tanenbaum Research Institute (LTRI) sequencing facility. Raw sequencing reads in Fastq formats were mapped onto mouse genome (mm9) using Tophat 1.4.1 and the RPKMs (reads per kilobase of exon model per million mapped reads) were calculated using a customized script. RNA-seq data are presented in Supplementary Table 1. Combined fold change presented in Extended Data Fig. 3d and Supplementary Table 1 was calculated using the following formula: combination fold change = $\log_2[(Yap^{\Delta/Δ}/Yap^{+/Δ})/(Dox+/Dox-)]$. R, Cluster 3.0 and Java TreeView were used for data visualization.

Organoids. Gut organoids were cultured according to a previously described protocol established by Sato and Clevers⁷. Briefly, crypts were harvested by incubating opened small intestines in PBS containing 2 mM EDTA. The epithelium was released by vigorous shaking and crypts separated using a 70 μ m cell strainer. Crypts were seeded in growth factor reduced Matrigel (BD Biosciences) and grown in Advanced DMEM/F12 (Invitrogen) supplemented with 2 mM GlutaMax (Invitrogen), 100 U ml⁻¹ Penicillin/100 μ g ml⁻¹ Streptomycin (Invitrogen), N2

Supplement (Invitrogen), B-27 Supplement (Invitrogen Cat), mouse recombinant Egf (R&D Systems), 100 ng ml⁻¹ mouse recombinant Noggin (Peprotech), 150 ng ml⁻¹ human Rsp1 (R&D Systems). Apc-deficient organoids were harvested from *Yap^{fl/+};Apc^{fl/fl};villin-creERT*, *Yap^{fl/fl};Apc^{fl/fl};villin-creERT* or *YapTg;Apc^{fl/fl};villin-creERT* mice injected with tamoxifen and seeded 48 h later in basal growth medium without Egf, Rsp1 or Noggin. To induce Yap expression in *YapTg* organoids, 1.5 µg ml⁻¹ doxycycline was added to the culture medium on day 0. Egf (R&D Systems, Cat. no. AF2028), Areg (R&D Systems, Cat. no. AF989) and Ereg (R&D Systems, Cat. no. 1068-EP-050) were added to the culture medium at a final concentration of 0.5 µg ml⁻¹. The following inhibitors were used: PD153053 (0.5 µM, Tocris Bioscience), U0126 (10 µM, Merck Millipore). To examine pErk1/2 levels, organoids were harvested at day 2 in cold PBS containing 5 mM EDTA, 1 mM NaVO₄, 1.5 mM NaF and protease inhibitors. Organoids were incubated at 4°C for 30 min to dissolve Matrigel and then lysed in TNTE buffer (50 mM Tris/HCl pH 7.6, 150 mM NaCl, 0.5% Triton X-100, 1 mM EDTA) containing standard protease and phosphatase inhibitors. Protein

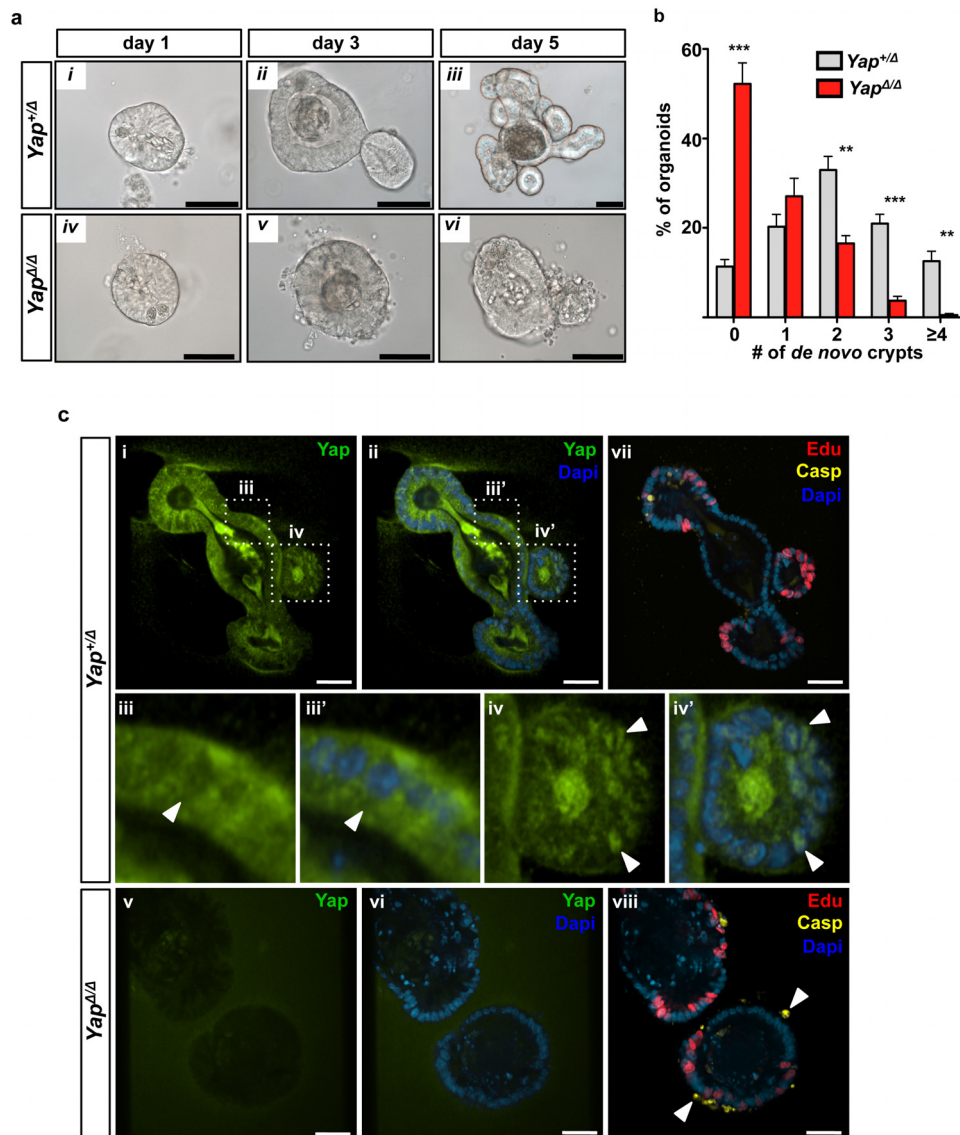
concentrations were measured and samples were subjected to SDS-PAGE. Total RNA was extracted by removing culture medium and directly lysing organoids in wells using RTL buffer of the RNeasy Mini Kit (Qiagen). RNA was purified using columns and genomic DNA was removed by treatment with RNase-Free DNase (Qiagen).

30. Reginensi, A. *et al.* Yap- and Cdc42-dependent nephrogenesis and morphogenesis during mouse kidney development. *PLoS Genet.* **9**, e1003380 (2013).
31. Gregorieff, A. *et al.* Expression pattern of Wnt signaling components in the adult intestine. *Gastroenterology* **129**, 626–638 (2005).
32. Lauter, G., Soll, I. & Hauptmann, G. Multicolor fluorescent *in situ* hybridization to define abutting and overlapping gene expression in the embryonic zebrafish brain. *Neural Dev.* **6**, 10 (2011).
33. Kosman, D. *et al.* Multiplex detection of RNA expression in *Drosophila* embryos. *Science* **305**, 846 (2004).
34. Vize, P. D., McCoy, K. E. & Zhou, X. Multichannel wholemount fluorescent and fluorescent/chromogenic *in situ* hybridization in *Xenopus* embryos. *Nature Protocols* **4**, 975–983 (2009).



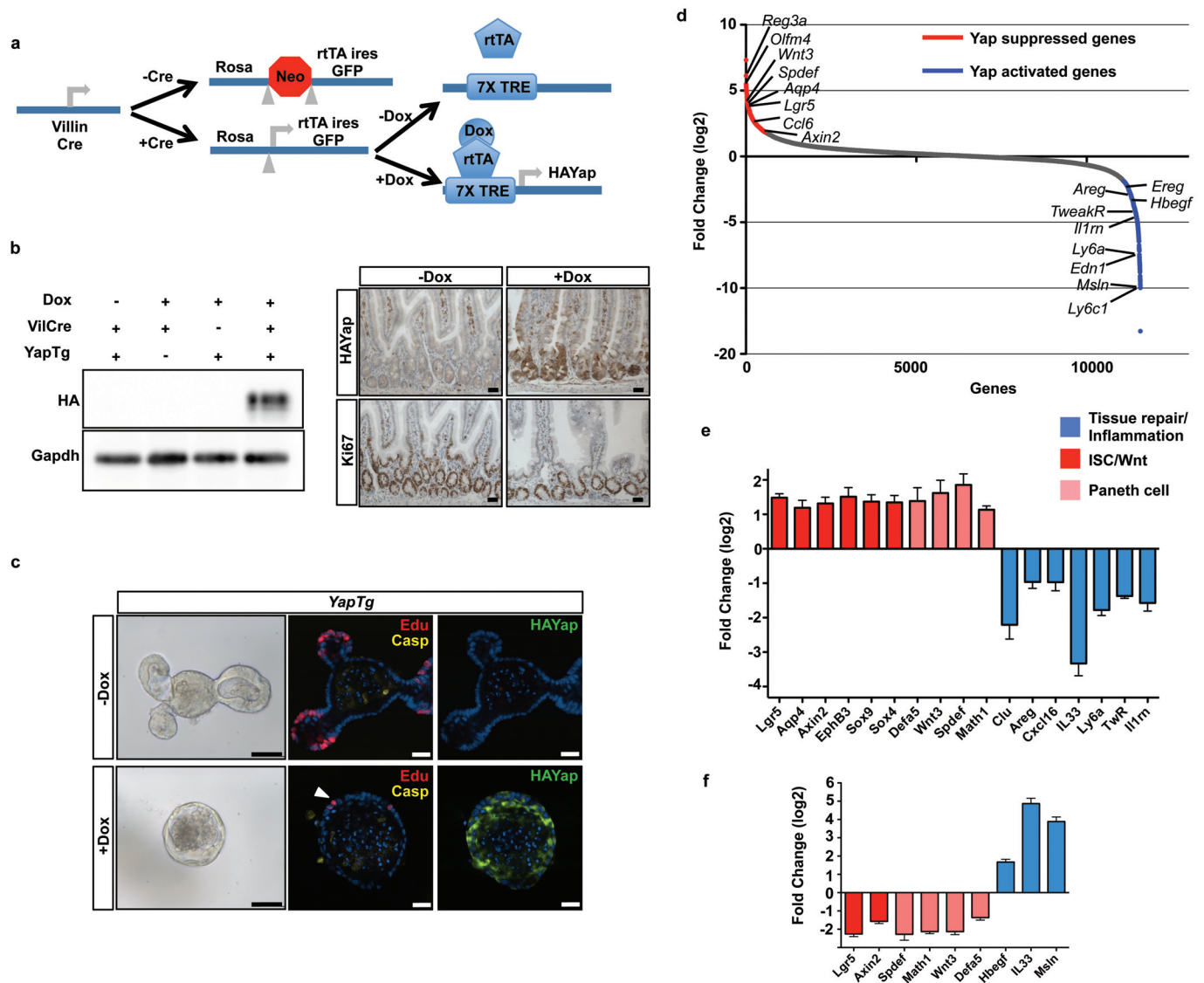
Extended Data Figure 1 | Analysis of late regenerative responses in Yap-deficient crypts after irradiation. **a**, Staining of untreated *Yap*^{+/-Δ} and *Yap*^{Δ/Δ} intestines with Yap/Taz antibodies. Filled arrowheads point to crypt base columnar cell (CBC) stem cells and open arrowheads indicate Paneth cells. **b**, Comparison of crypt proliferation in *Yap*^{+/-Δ} and *Yap*^{Δ/Δ} mice at 3, 4 and 6 dpi (10 Gy) by staining representative sections with Ki67 antibodies. Sections in bottom panels were immunostained with anti-Yap/Taz antibodies to confirm the absence in Yap and Taz expression in *Yap*^{Δ/Δ} crypts at 6 dpi. **c**, Stainings of control *Yap*^{+/-Δ}; *Taz*^{+/-Δ} versus *Yap*^{Δ/Δ}; *Taz*^{Δ/Δ} mice at 6 dpi (10 Gy) with Ki67 and Yap/Taz antibodies. **d**, Mosaic analysis of Yap in the late regenerative response. Pairs of consecutive sections from mice displaying

mosaic intestinal expression of Yap at 6 days post-irradiation (10 Gy) were stained for Yap (left panel) and Ki67 (right panel). Open arrowheads in consecutive sections represent Yap-null crypts and filled arrowheads point to Yap-positive crypts. Images in **b** and **c** are representative of at least three stainings performed on tissues derived from separate mice. **e, f**, Analysis of apoptotic and mitotic cells in *Yap*^{+/-Δ} and *Yap*^{Δ/Δ} intestines at 1 dpi (12 Gy). Representative stainings of anti-active caspase 3 and BrdU incorporation are shown to the left. Bar graphs represent percentage of caspase 3⁺ cells and BrdU⁺ cells within the crypt epithelium scored from at least 4 individual mice per genotype. Error bars indicate s.e.m.; *n* = 30 (*n* represents the total number of sections scored per genotype); ****P* < 0.0001. Scale bars, 70 μm.



Extended Data Figure 2 | Yap localization and function in organoid cultures. **a, b,** Crypts from *Yap^{+/-}* and *Yap^{Δ/Δ}* mice were harvested and cultured under standard conditions (see Methods) for the indicated times. Panel **b** depicts the percentage of organoids showing 0, 1, 2, 3 or ≥ 4 *de novo* crypts at 4 days (error bars indicate s.e.m.; $n = 7$ (n represents the number of separate cultures per genotype per mouse; *** $P = 0.0006$, ** $P < 0.0021$)).

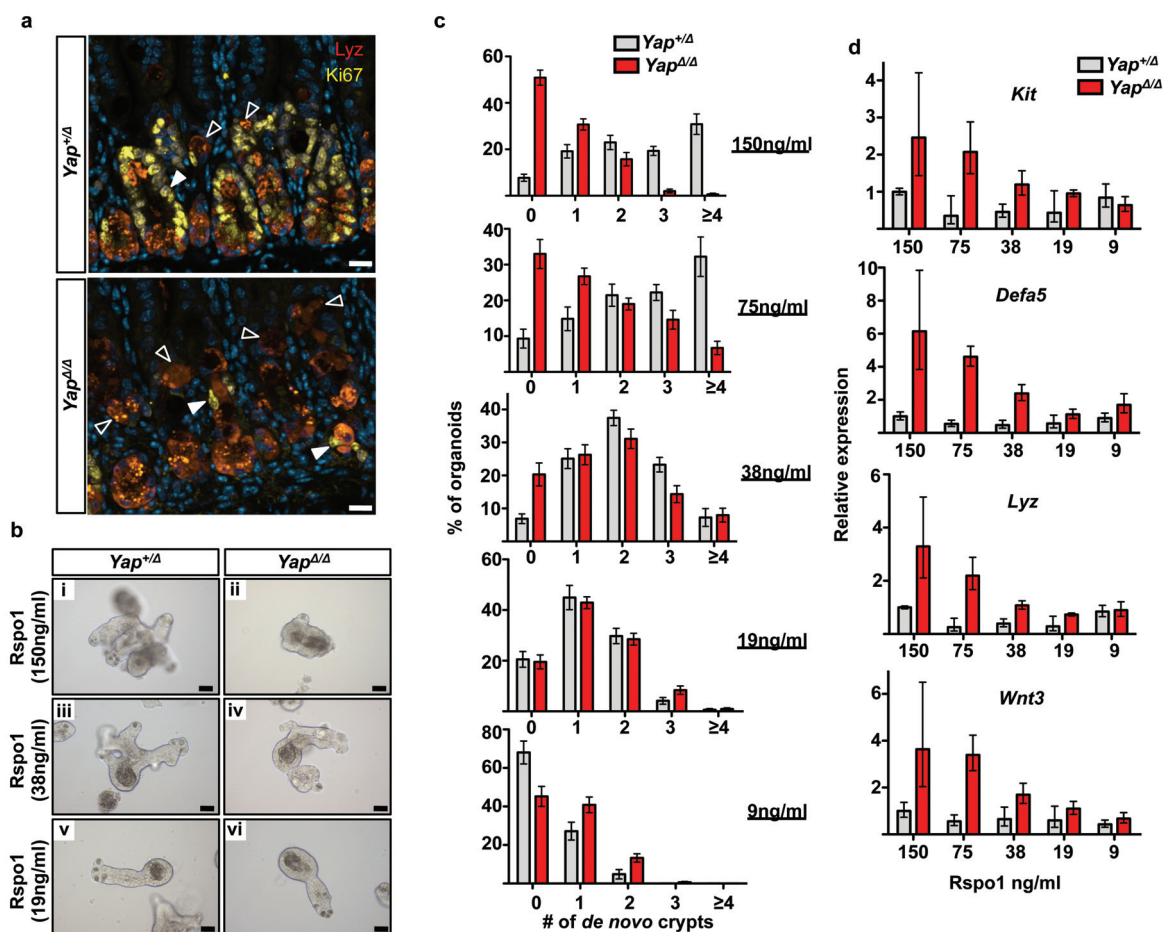
c, Proliferation and apoptosis in *Yap^{+/-}* and *Yap^{Δ/Δ}* organoids grown for 3 days were evaluated by examining incorporation of Edu (red) and active caspase 3 (yellow), respectively. Endogenous Yap expression is shown in green. Panels *iii* and *iii'* show cytoplasmic localization of Yap. Panels *iv* and *iv'* show nuclear accumulation of Yap in forming crypts. Arrowheads indicate increased apoptotic cells in *Yap^{Δ/Δ}* organoids. Scale bars, 70 μm .



Extended Data Figure 3 | Identification of Yap regulated genes.

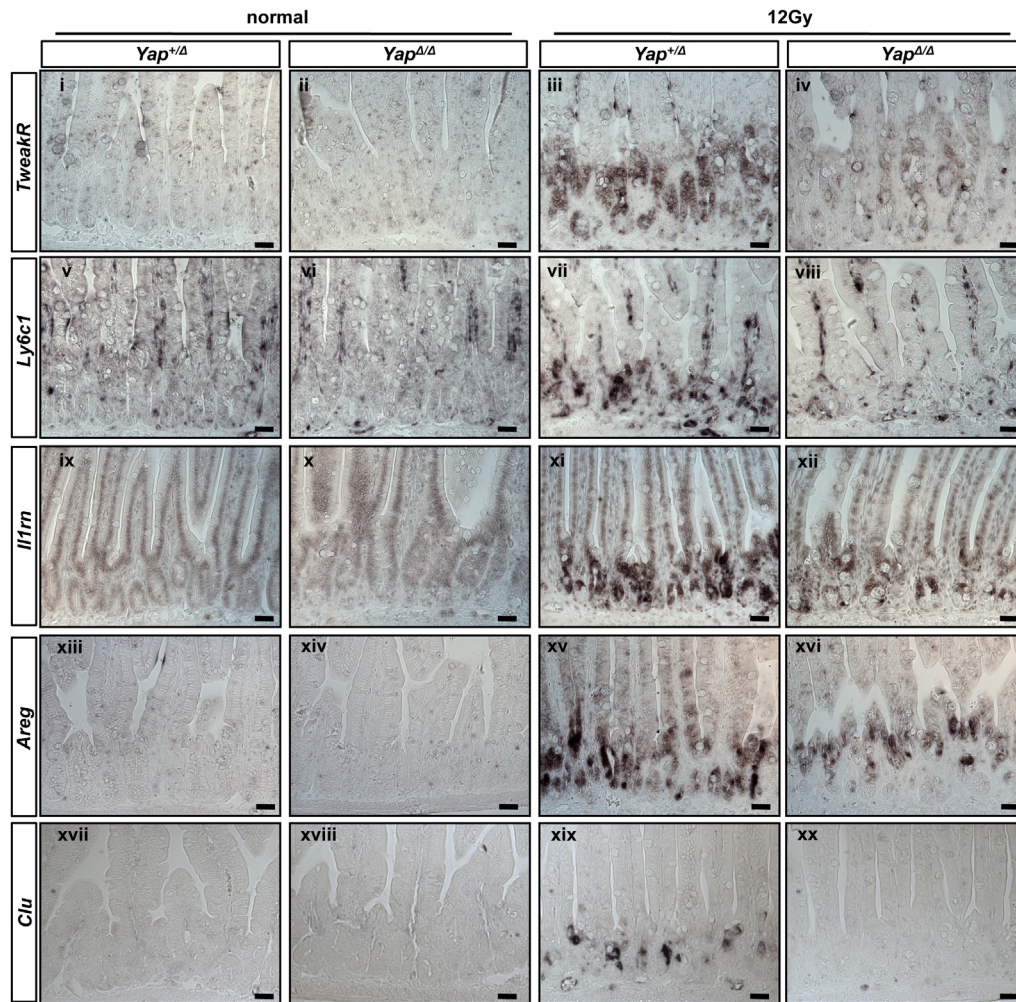
a, Schematic representation of *YapTg* mice. Induction of HA-Yap was achieved by intercrossing *villin-cre* or *villin-creERT2* mice with *Rosa26-lox-STOP-lox-rtta-IRES-EGFP* mice (see Methods). **b**, Analysis of HA-Yap protein expression in the intestinal epithelium. Intestinal crypts were isolated, lysed and subjected to SDS-PAGE (left panel). Expression of HA-Yap is only detected in transgenic mice in the presence doxycycline. Immunohistochemistry staining using anti-HA and Ki67 antibodies in untreated small intestine of *YapTg* mice (right panel). **c**, Analysis of Yap overexpression in organoid cultures. Crypts from *YapTg* intestines were seeded and induced with doxycycline. Representative organoids cultured for 3 days are shown as bright-field images or Edu (red) and caspase 3 (yellow) stainings. Yap transgene

expression was detected by anti-HA staining (green) ($n = 5$). Arrowhead indicates diminished Edu incorporation in doxycycline-induced organoids. Scale bars, 70 μm . **d**, Identification of relative expression of Yap-regulated genes by RNA-seq analysis are shown as rank order plots comparing control and *Yap*^{+/-} and *Yap*^{Δ/Δ} organoids isolated at day 1, as well as doxycycline treated and untreated *YapTg* organoids at day 1 of culture: combined fold change = $\log_2 [(Yap^{\Delta/Δ}/Yap^{+/-})/YapTg(Dox+/Dox-)]$. **e**, **f**, qPCR analysis of selected Yap-regulated genes comparing fold change between *Yap*^{+/-} and *Yap*^{Δ/Δ} or doxycycline-treated and untreated *YapTg* organoids at day 1, respectively. Error bars indicate s.e.m.; $n > 3$ (n represents the number of independent organoid cultures per genotype per mouse analysed for each gene).



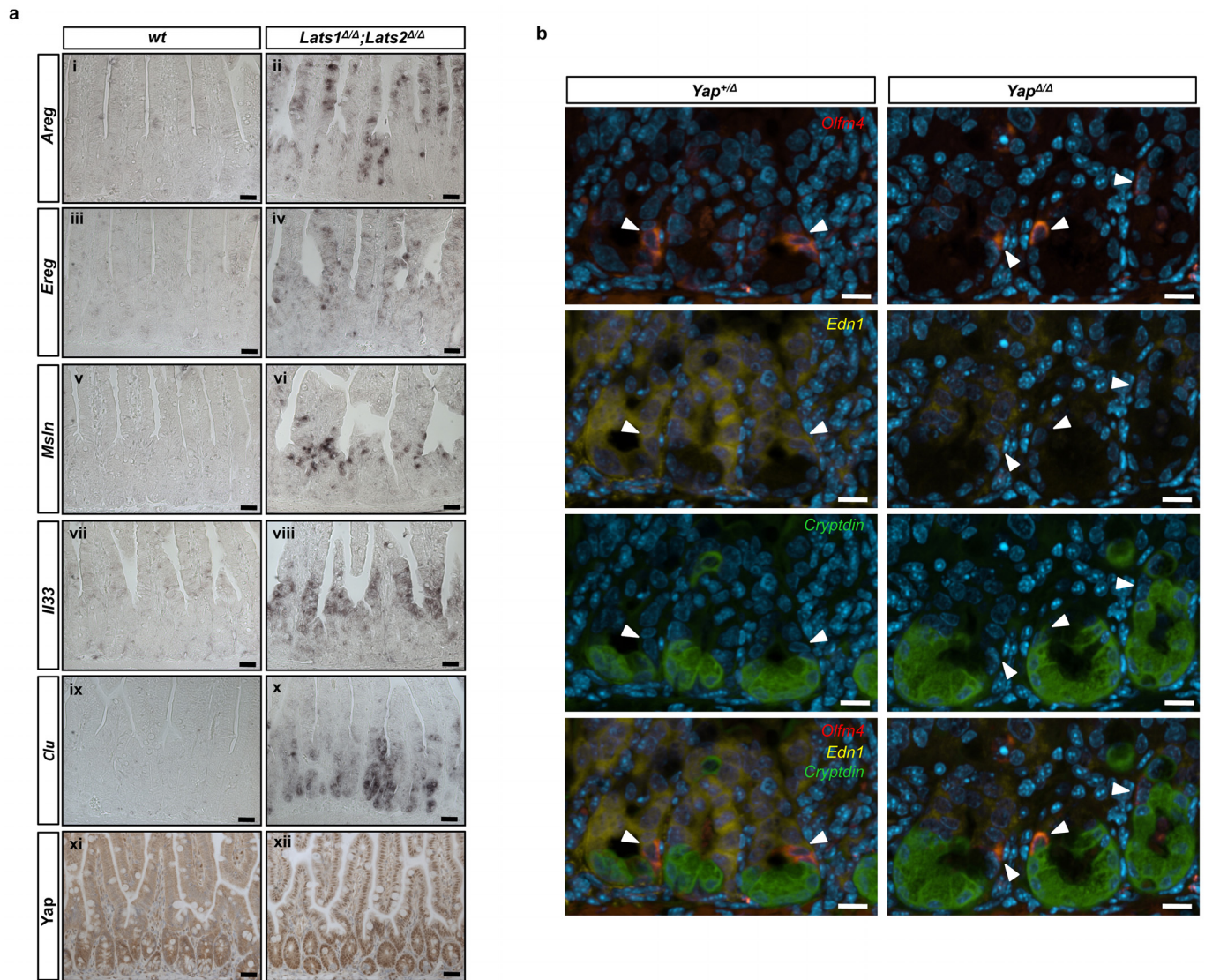
Extended Data Figure 4 | Lowering Rspo1 levels rescues growth of Yap mutant organoids. **a**, Fluorescence immunostaining of *Yap*^{+Δ} and *Yap*^{Δ/Δ} mice at 3 dpi to detect Ki67⁺ cells (yellow) and lysozyme positive (Lyz⁺) Paneth cells (red). **b**, Representative images of *Yap*^{+Δ} (i, iii, v) and *Yap*^{Δ/Δ} (ii, iv, vi) organoids cultured in reducing Rspo1 concentrations for 3 days. Scale bars: **a**, 35 μm and **b**, 70 μm. **c**, Quantification of the percentage of organoids displaying 0, 1, 2, 3 or ≥4 *de novo* crypts after 3 days in culture using indicated

concentrations of Rspo1. Error bars indicate s.e.m.; *n* = 4 (*n* represents number of organoid cultures analysed per genotype per mouse). **d**, Relative expression of Paneth cell markers (*Kit*, *Lyz*, *Wnt3*, *Defa5*) was evaluated by qPCR. Graphs show representative results of 3 independent organoid cultures per genotype. Error bars represent minimal and maximal range in fold differences derived from the standard deviation in Ct values; *n* = 3 (*n* represents the number of technical repeats).



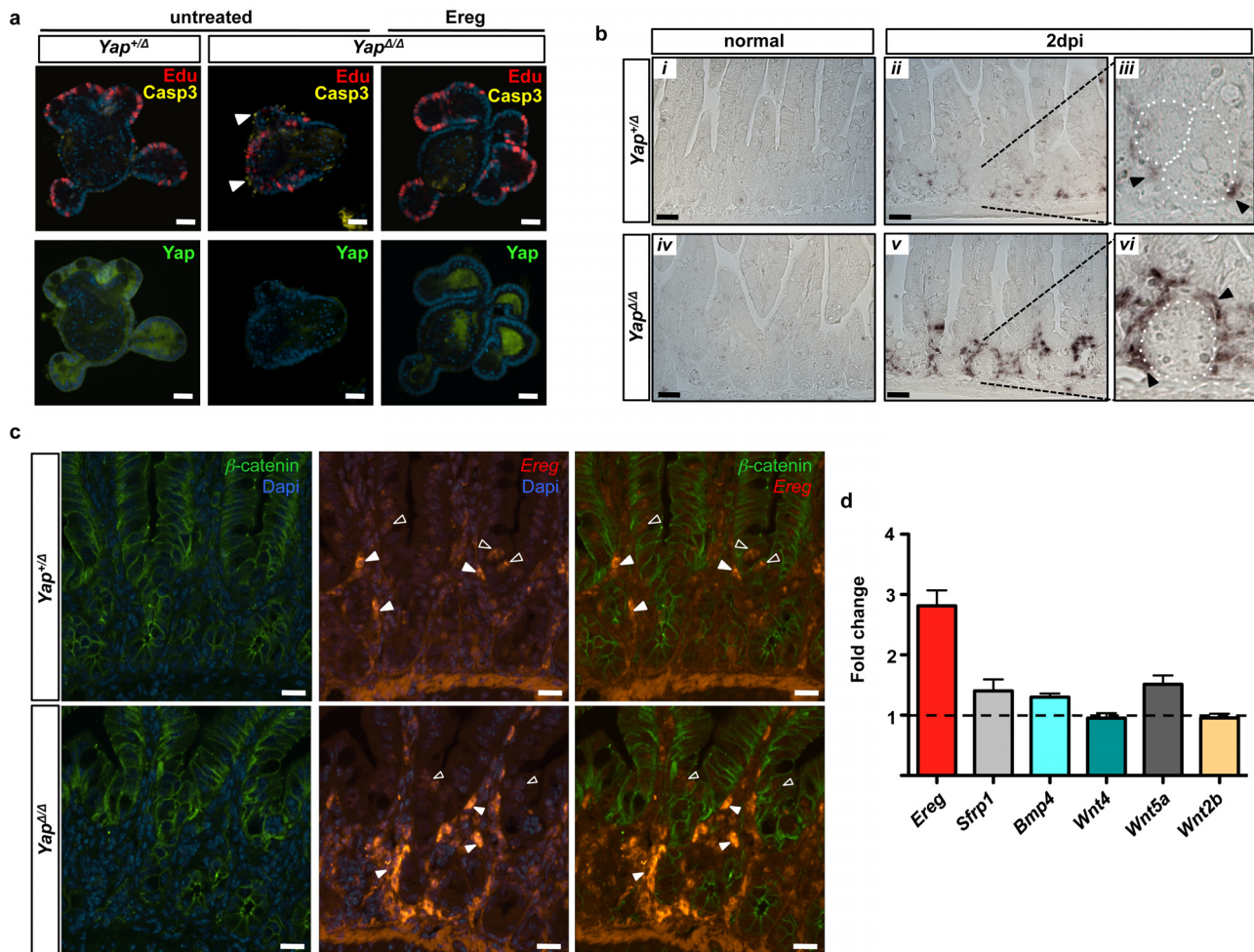
Extended Data Figure 5 | Yap-dependent expression of pro-regenerative genes in the intestinal epithelium after irradiation. ISH on untreated and irradiated (2 dpi, 12 Gy) *Yap*^{+/Δ} and *Yap*^{Δ/Δ} intestines using specific probes for

TweakR (i–iv), *Ly6c1* (v–viii), *Il1rn* (ix–xii), *Areg* (xiii–xvi) and *Clu* (xvii–xx). Images are representative of at least two stainings per gene performed on tissues derived from separate mice. Scale bars, 70 μm.



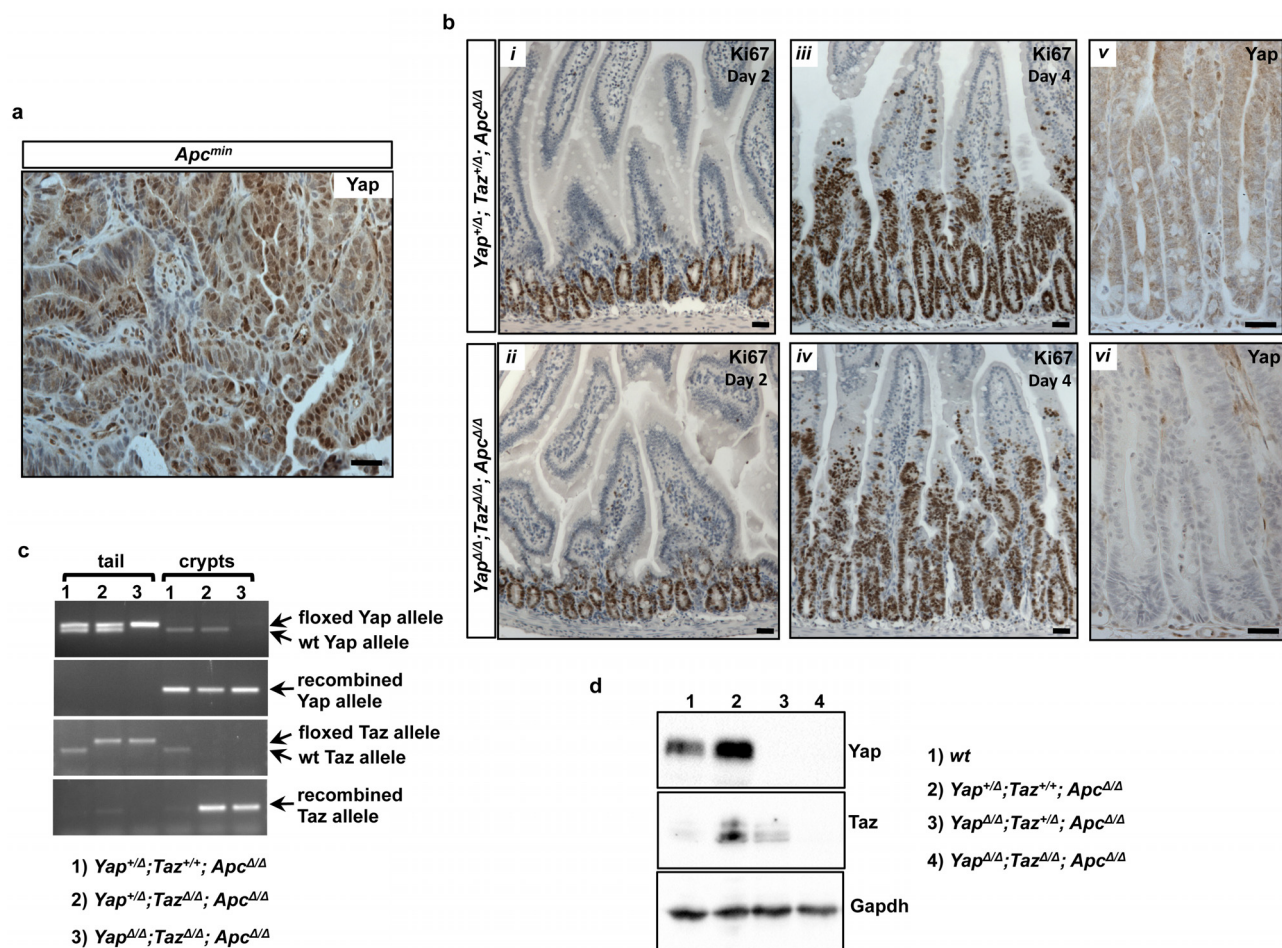
Extended Data Figure 6 | The Yap-dependent regenerative program is activated in *Lgr5*⁺ ISCs. **a**, ISH showing induction of *Areg* (i–ii), *Ereg* (iii–iv), *Msln* (v–vi), *Il33* (vii–viii), *Clu* (ix–x) and *Yap* protein expression (xi–xii) in intestinal epithelium of *Lats1^{Δ/Δ};Lats2^{Δ/Δ}* mice. **b**, Fluorescence ISH to detect *Edn1*, *Olfm4* and cryptdin1 expression in irradiated (2 dpi; 12 Gy) *Yap^{+/-}* and

Yap^{Δ/Δ} intestines ($n = 3$). *Edn1* expression is detected in ISCs post-irradiation in a Yap-dependent manner. Arrowheads point to location of *Olfm4*⁺ ISCs. Images are representative of three stainings per gene performed on tissues derived from separate mice. Scale bars: **a**, 70 μm ; **b**, 35 μm .



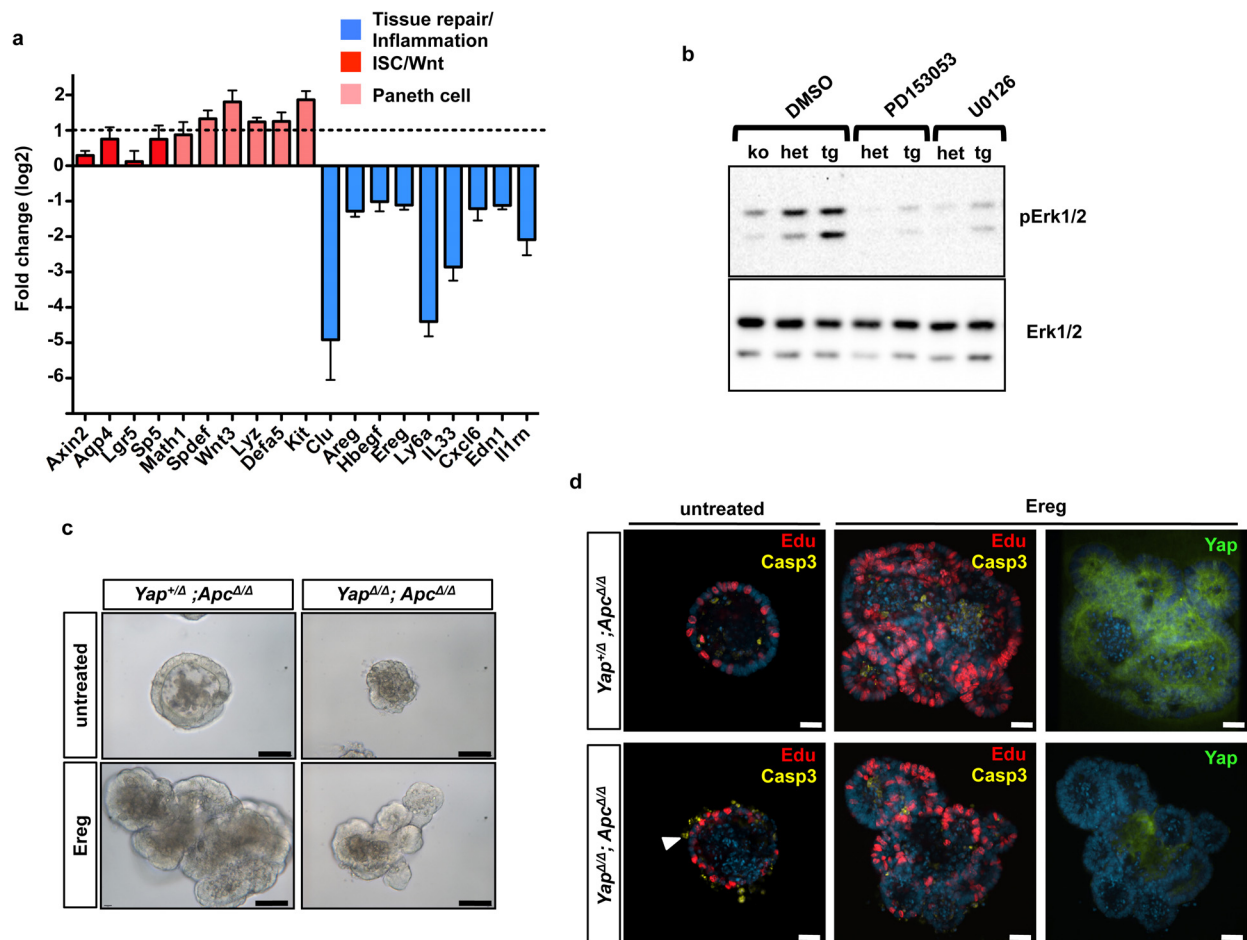
Extended Data Figure 7 | Stimulation of Yap-independent growth via stromally derived Ereg expression. **a**, *Yap*^{+/ Δ} and *Yap* ^{Δ / Δ} organoids were grown for 3 days in standard growth media and supplemented with 0.5 $\mu\text{g ml}^{-1}$ of recombinant Ereg. Panels in the top row show Edu incorporation and active caspase 3 stainings in organoid cultures. Proliferation and apoptosis status in Ereg-stimulated *Yap* ^{Δ / Δ} organoids are comparable to control organoids. Arrowhead points to apoptotic cells in Yap-deficient organoids. Panels in the bottom row show endogenous Yap expression and confirm that Ereg-stimulated *Yap* ^{Δ / Δ} organoids are Yap deficient. Green signal in *Yap* ^{Δ / Δ} organoids is non-specific staining of cellular debris in lumen. **b**, ISH to monitor *Ereg* expression in untreated and irradiated (2 dpi, 12 Gy) *Yap*^{+/ Δ} and

Yap ^{Δ / Δ} mice. In iii and vi, dotted lines demarcate crypt boundaries. **c**, Ereg was detected by fluorescence ISH (red) and epithelia highlighted by counterstaining for β -catenin protein (green). Open arrowheads point to examples of Ereg expression in the epithelium, and filled arrowheads indicate stromal cells. Note Ereg expression in certain cells of the regenerating epithelium. All images are representative of three stainings performed on tissues derived from separate mice. Scale bars (**a–c**), 70 μm . **d**, qPCR analysis of stromally derived factors *Ereg*, *Sfrp1*, *Bmp4*, *Wnt2b*, *Wnt4*, *Wnt5a* from samples of whole intestines of irradiated (2 dpi; 10 Gy) *Yap*^{+/ Δ} and *Yap* ^{Δ / Δ} mice. Error bars indicate s.e.m.; $n = 7$ (n represents number of independent mice analysed per genotype).



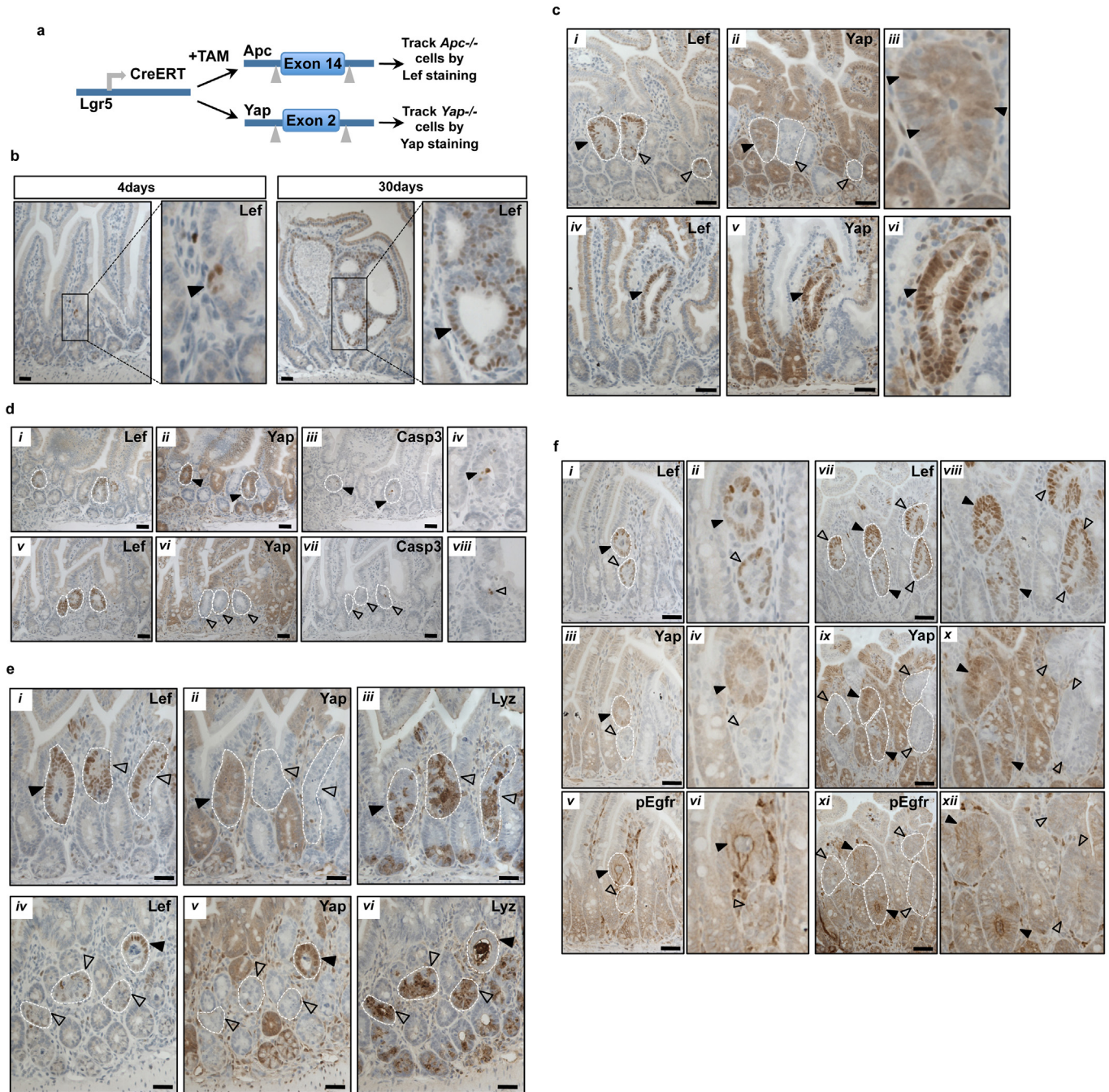
Extended Data Figure 8 | Role of Yap in *Apc* mutant cells. **a**, $Apc^{Min/+}$ adenomas display high levels of nuclear Yap. **b**, Comparison of crypt proliferation in $Yap^{+/Δ};Taz^{+/Δ};Apc^{Δ/Δ}$ and $Yap^{Δ/Δ};Taz^{Δ/Δ};Apc^{Δ/Δ}$ mice at 2 and 4 days after tamoxifen injection by staining representative sections with Ki67 antibodies (panels i–iv). Sections in panels v and vi were immunostained

with anti-Yap/Taz antibodies. Images are representative of three stainings performed on tissues derived from separate mice. Scale bars, 70 μ m. **c**, **d**, Genotyping of tail and crypt DNA and western blot analysis from indicated mice at 4 days after tamoxifen injection to confirm deletion of Yap and Taz.



Extended Data Figure 9 | Role of Yap- and Egfr-dependent signalling in *Apc* mutant cells. **a**, qPCR analysis of selected Yap regulated genes comparing fold change between *Apc* mutant *Yap^{+/-}* and *Yap^{Δ/Δ}* organoids. Error bars indicate s.e.m.; $n > 4$ (n represents the number of independent organoid cultures per genotype per mouse analysed for each gene). **b**, Western blot analysis showing expression of pErk1/2 in *Yap^{+/-};Apc^{Δ/Δ}* (het), *Yap^{Δ/Δ};Apc^{Δ/Δ}* (KO) and doxycycline-treated *YapTg;Apc^{Δ/Δ}* organoids (Tg). Effects of PD153035 and U0126 treatment on pErk1/2 levels are also shown ($n = 3$; n represents the number of separate organoid cultures analysed per genotype).

c, d, *Yap^{+/-};Apc^{Δ/Δ}* and *Yap^{Δ/Δ};Apc^{Δ/Δ}* mice were injected with tamoxifen and crypts harvested 48 h later. Organoids were grown in medium lacking Rsp1, Noggin and Egf and supplemented with $0.5 \mu\text{g ml}^{-1}$ of recombinant Ereg. Immunofluorescence shows that Ereg treatment strongly enhances Edu incorporation in both control and *Yap^{Δ/Δ};Apc^{Δ/Δ}* (middle panels). Yap stainings (right panels) show that Ereg-stimulated *Yap^{Δ/Δ};Apc^{Δ/Δ}* organoids lack endogenous Yap. Images are representative of at least three cultures derived from individual mice. Scale bars, 70 μm .



Extended Data Figure 10 | Role of Yap in *Apc* mutant tumour initiating cells. **a**, Experimental procedure to target Yap in *Apc* mutant *Lgr5*⁺ ISCs. **b**, Tracing of *Apc*^{Δ/Δ} cells by staining for the Wnt target gene Lef at 4 and 30 days after tamoxifen injection (dpi) of *Yap;Apc*^{ALgr5-cre} mice. Note Lef staining is absent in surrounding wild-type crypts. **c**, Consecutive sections from *Yap;Apc*^{ALgr5-cre} mice at 10 dpi demonstrate both Yap negative (open arrowheads) and Yap positive (filled arrowheads) Lef⁺ foci (panels i–iii). Arrowheads in panel iii indicate occasional nuclear Yap staining in early Lef⁺ foci. In Lef⁺ foci displaying aberrant crypt morphology, Yap is strongly

nuclear (panels iv–vi). **d**, Consecutive sections from *Yap;Apc*^{ALgr5-cre} mice at 10 dpi showing the levels of caspase 3⁺ apoptotic cells in both Yap-positive (filled arrowheads, panels i–iv) and Yap-negative (open arrowheads, panels v–viii) Lef⁺ foci. **e**, Two sets of consecutive sections (i–iii and iv–vi) showing Lef, Yap and Lyz staining of representative Lef⁺ foci. **f**, Two sets of consecutive sections (i–vi and vii–xii) showing Lef, Yap and phospho-Egfr staining of tamoxifen-induced *Yap;Apc*^{ALgr5-cre} mice. Filled and open arrowheads indicate Yap-positive and -negative Lef⁺ foci, respectively and panels ii, iv, vi, viii, x and xii are enlargements of adjacent panels. Scale bars, 70 μm.

Bacteriocin production augments niche competition by enterococci in the mammalian gastrointestinal tract

Sushma Kommineni^{1,3}, Daniel J. Bretl³, Vy Lam¹, Rajrupa Chakraborty^{1,3}, Michael Hayward¹, Pippa Simpson², Yumei Cao², Pavlos Bousounis¹, Christopher J. Kristich³ & Nita H. Salzman^{1,3}

Enterococcus faecalis is both a common commensal of the human gastrointestinal tract and a leading cause of hospital-acquired infections¹. Systemic infections with multidrug-resistant enterococci occur subsequent to gastrointestinal colonization². Preventing colonization by multidrug-resistant *E. faecalis* could therefore be a valuable approach towards limiting infection. However, little is known about the mechanisms *E. faecalis* uses to colonize and compete for stable gastrointestinal niches. Pheromone-responsive conjugative plasmids encoding bacteriocins are common among enterococcal strains³ and could modulate niche competition among enterococci or between enterococci and the intestinal microbiota. We developed a model of colonization of the mouse gut with *E. faecalis*, without disrupting the microbiota, to evaluate the role of the conjugative plasmid pPD1 expressing bacteriocin 21 (ref. 4) in enterococcal colonization. Here we show that *E. faecalis* harbouring pPD1 replaces indigenous enterococci and outcompetes *E. faecalis* lacking pPD1. Furthermore, in the intestine, pPD1 is transferred to other *E. faecalis* strains by conjugation, enhancing their survival. Colonization with an *E. faecalis* strain carrying a conjugation-defective pPD1 mutant subsequently resulted in clearance of vancomycin-resistant enterococci, without plasmid transfer. Therefore, bacteriocin expression by commensal bacteria can influence niche competition in the gastrointestinal tract, and bacteriocins, delivered by commensals that occupy a precise intestinal bacterial niche, may be an effective therapeutic approach to specifically eliminate intestinal colonization by multidrug-resistant bacteria, without profound disruption of the indigenous microbiota.

The mammalian host is colonized by 10–100 trillion microbes that live in a predominantly symbiotic relationship with their host^{5,6}. Invasion by some of these symbionts can cause serious disease when homeostasis is disrupted⁷. One such symbiont is *E. faecalis*, a Gram-positive member of the gut microbiota of a wide range of mammals, including humans⁸. Although enterococci are usually not pathogenic, they can cause significant disease in immune-compromised individuals. Treating enterococcal infections is challenging owing to their intrinsic and acquired resistance to a wide range of antibiotics^{9,10}. Prior antibiotic therapy is a well-known risk factor for enterococcal infection^{9,11}. Antibiotic disruption of the intestinal biota enables resistant enterococci to proliferate profusely in the gastrointestinal tract¹² and invade the host¹¹. Reducing colonization of the gastrointestinal tract by antibiotic-resistant enterococci could therefore represent a promising approach for preventing enterococcal infections. However, our mechanistic understanding of enterococcal gastrointestinal colonization is limited. Most previous studies employed antibiotic disruption of the intestinal microbiota before enterococcal challenge, resulting in invasion rather than colonization. Our model establishes long-term colonization of mice with a marked strain of *E. faecalis* without antibiotic

pretreatment, allowing us to study *E. faecalis* dynamics in an unperturbed intestinal environment.

We established long-term colonization of FVB mice by prolonged feeding with a rifampicin-resistant *E. faecalis* strain (EF_r), CK135. After withdrawal of EF_r from the drinking water, mice maintained persistent colonization of the gastrointestinal tract and shedding of EF_r in faeces for more than eleven weeks (data shown for first four weeks, Fig. 1a, b). In contrast, mice gavaged with ~10⁹ colony-forming units (CFU) of EF_r demonstrated short-lived colonization (2–3 days), with significant mouse-to-mouse variability. Although the specific factors that determine bacterial gastrointestinal colonization fitness are unclear, the success of our model depends on both the nature and the route of delivery and offers a means of investigating the underlying mechanisms. Similar levels and patterns of EF_r colonization were observed in C57Bl/6 mice and with multiple lineages of *E. faecalis* (Extended Data Fig. 1a, b), suggesting that this approach for mouse gastrointestinal colonization by *E. faecalis* is generalizable.

We sought to understand the effect of bacteriocins—antimicrobial peptides often encoded on pheromone-responsive plasmids—on enterococcal colonization dynamics of the gastrointestinal tract. Bacteriocin production is a tool bacteria can use to enhance the stability of their communities¹³, by outcompeting closely related bacterial species and establishing a stable niche for the producing strain^{14,15}. Many lactic acid bacteria produce bacteriocins^{15–18}. The enterococcal bacteriocin, bacteriocin 21 (bac-21), is encoded on the sex-pheromone-responsive conjugative plasmid pPD1 (ref. 19) and is identical in nucleotide sequence to another well-characterized enterococcal bacteriocin, AS-48 (ref. 20) (Supplementary Table 1a, b). The prevalence of pPD1, first isolated from an enterococcal strain of human gingival origin, has not been described. Only the pheromone-responsive and bacteriocin operons (*bac*) of pPD1 have previously been sequenced^{14,19}. Complete sequencing and annotation of pPD1 revealed a 57,732-base-pair (bp) plasmid with a total of 59 open reading frames (ORFs) (Extended Data Fig. 2 and Supplementary Table 2). The pPD1 *bac* operon contains nine genes, designated from *bacA* to *bacI*¹⁹, of which *bacA* encodes a 105-amino-acid precursor of bac-21 (ref. 19). The genes *bacB* to *bacI* are thought to be involved in processing, modification and secretion of the mature 70-amino-acid bacteriocin and providing immunity to the bacteriocin producer¹⁹.

Colonization experiments with an *E. faecalis* strain harbouring pPD1 (EF_r + pPD1) revealed that EF_r + pPD1 was significantly more abundant than EF_r in faeces and throughout the gastrointestinal tract, suggesting more effective colonization (Fig. 1a, b). To test whether pPD1-encoded bac-21 drives enhanced enterococcal colonization, we introduced an in-frame deletion of *bacAB* into pPD1 (EF_r + pPD1::Δ*bacAB*). EF_r + pPD1::Δ*bacAB* did not exhibit a growth defect *in vitro* and lacked bacteriocin activity (Extended Data Fig. 3a). Mice were colonized with EF_r + pPD1, EF_r + pPD1::Δ*bacAB* or EF_r,

¹Division of Gastroenterology, Department of Pediatrics, Medical College of Wisconsin, Milwaukee, Wisconsin 53226, USA. ²Division of Quantitative Health Sciences, Department of Pediatrics, Medical College of Wisconsin, Milwaukee, Wisconsin 53226, USA. ³Department of Microbiology and Molecular Genetics, Medical College of Wisconsin, Milwaukee, Wisconsin 53226, USA.

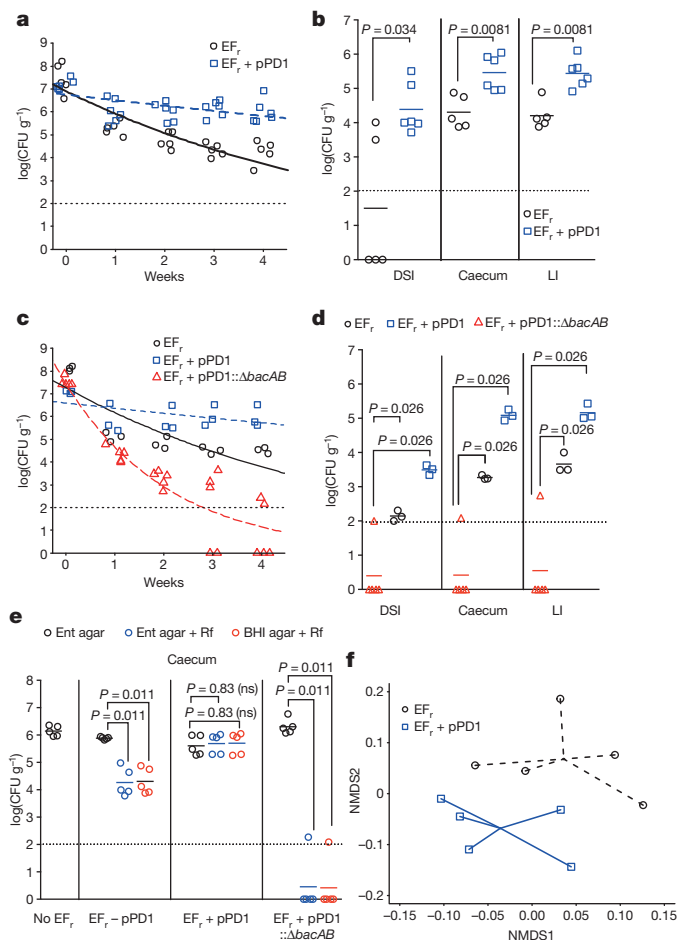


Figure 1 | pPD1 enhances *E. faecalis* competition for an intestinal niche.

a, b, Mice were colonized by EF_r ($n = 5$) or $\text{EF}_r + \text{pPD1}$ ($n = 6$), which were enumerated weekly from faeces (**a**) and at week 4 from each segment of the gastrointestinal tract (**b**) (distal small intestine (DSI), caecum and large intestine (LI)). **c, d**, Mice were colonized with $\text{EF}_r + \text{pPD1}::\Delta\text{bacAB}$ ($n = 5$), EF_r ($n = 3$) or $\text{EF}_r + \text{pPD1}$ ($n = 3$). Faecal abundance of *E. faecalis* was determined weekly (**c**) and in the gastrointestinal tract at week 4 (**d**). **e**, Mice ($n = 5$ per group) were colonized with EF_r , $\text{EF}_r + \text{pPD1}$, $\text{EF}_r + \text{pPD1}::\Delta\text{bacAB}$ or sterile drinking water; at week 4, abundance of total enterococci (indigenous and laboratory strains) and rifampicin-resistant laboratory strains of *E. faecalis* was enumerated in each segment of the gastrointestinal tract. **f**, Microbiome analysis. Ordination of *E. faecalis* ($n = 5$ mice) and $\text{EF}_r + \text{pPD1}$ ($n = 5$ mice) samples was separated by the Bray–Curtis beta diversity metric (Adonis $P = 0.007$). Samples are connected to help visualize grouping. An exponential decay model was fitted to the data in **a** and **c**. In **a**, the rate of decay is significantly different between the two groups ($P < 0.0001$) and in **c** between $\text{EF}_r + \text{pPD1}::\Delta\text{bacAB}$ and EF_r ($P < 0.0001$), as well as between $\text{EF}_r + \text{pPD1}::\Delta\text{bacAB}$ and $\text{EF}_r + \text{pPD1}$ ($P < 0.0001$). Horizontal lines in **b**, **d** and **e** indicate geometric means. Each symbol represents an individual animal; data are representative of six (**a, b**) and two (**e**) biologically independent experiments. Data are from one experiment in **c, d** and **f**. Black dotted lines (all panels) indicate the limit of detection at 100 CFU per g faeces. BHI, brain-heart infusion medium; Ent, Enterococcus agar; NMDS, non-metric multidimensional scaling; ns, not significant; Rf, rifampicin.

$\text{EF}_r + \text{pPD1}::\Delta\text{bacAB}$ showed no colonization advantage over EF_r and demonstrated impaired long-term persistence in the gastrointestinal tract compared to $\text{EF}_r + \text{pPD1}$ and EF_r (Fig. 1c, d), possibly owing to the excessive burden of maintaining a large non-functional plasmid. Alternatively, it is possible that unknown colonization factors expressed in the presence of an intact pPD1 are affected by *bacAB* deletion, resulting in loss of colonization ability.

Mice obtained from commercial vendors possess a variable indigenous intestinal enterococcal population (Fig. 1e and Extended Data

Fig. 3b–e). To determine whether exogenous colonization with EF_r strains replaced the indigenous enterococci to establish a stable niche in the gut, the indigenous enterococcal population was monitored after colonization with EF_r or $\text{EF}_r + \text{pPD1}$ (Extended Data Fig. 3b, c). EF_r accounted for approximately 5–10% of the total enterococci in the faeces and declined over time to 0.1–1% (Fig. 1e and Extended Data Fig. 3b–e). However, $\text{EF}_r + \text{pPD1}$ rapidly dominated the niche within one week and maintained this dominance over time (Extended Data Fig. 3b, c), suggesting that it was displacing or outcompeting the majority of the indigenous enterococci in the mouse gut. $\text{EF}_r + \text{pPD1}::\Delta\text{bacAB}$ could not compete in the gut as it failed to displace the indigenous enterococci (Fig. 1e and Extended Data Fig. 3b–e). Moreover, colonization of the gut with EF_r carrying another pheromone-inducible conjugative plasmid that lacks bacteriocins (pCF10) did not result in domination of the intestinal niche (Extended Data Fig. 4). Hence, bac-21 is likely to be responsible for the colonization advantage conferred by pPD1.

Although this result suggests that $\text{EF}_r + \text{pPD1}$ directly kills competing enterococci in the gut, bac-21 might also alter physical or nutritional niches by killing other competing bacterial species^{4,19}. To investigate this, high-throughput 16S ribosomal DNA (rDNA) pyrosequencing was used to assess the composition of the gut community in colonized mice. We found no significant differences in composition between the caeca of control mice and those colonized by EF_r , indicating that colonization of mice with laboratory strains of *E. faecalis* does not substantially alter intestinal microbial ecology (Extended Data Fig. 5a). The caecal composition of EF_r and $\text{EF}_r + \text{pPD1}$ were significantly different ($P = 0.007$) and clustered separately (Fig. 1f). However, only a member of the Gram-negative Deferribacteraceae, *Mucispirillum*, was significantly different between groups (Extended Data Fig. 5b). We suggest that this may be a secondary effect, as Gram-positive bacteriocins such as bac-21 have low efficacy against Gram-negative bacteria. We conclude that elimination of competing enterococcal strains is the primary mechanism by which $\text{EF}_r + \text{pPD1}$ enhances its colonization.

To further investigate bac-21-mediated enterococcal competition in the gut, a competitive colonization experiment using differentially marked *E. faecalis* strains was performed. Combinations of EF_r and plasmid-carrying spectinomycin-resistant *E. faecalis* ($\text{EF}_s + \text{pPD1}$) or EF_r and $\text{EF}_s + \text{pPD1}$ with an in-frame deletion ($\text{EF}_s + \text{pPD1}::\Delta\text{bacAB}$) were given to groups of mice, at indicated ratios (Fig. 2). At week 0, faecal shedding of EF_r , $\text{EF}_s + \text{pPD1}$ and $\text{EF}_s + \text{pPD1}::\Delta\text{bacAB}$ reflected the ratio of the bacteria given in the drinking water (Fig. 2). When $\text{EF}_s + \text{pPD1}$ was given in excess, it rapidly outcompeted EF_r (Fig. 2g). When EF_r was given in excess (Fig. 2a) or equal abundance to $\text{EF}_s + \text{pPD1}$ (Fig. 2d), it did not outcompete $\text{EF}_s + \text{pPD1}$, despite being able to persist at lower levels. Subsequent experiments indicated that the ability of EF_r to persist was due to conjugative transfer of pPD1 to the EF_r host, thus confirming the competitive advantage conferred by bac-21. To control for unanticipated effects of the specific antibiotic resistance marker on the survival advantage of *E. faecalis*, the reciprocal experiment using EF_s and $\text{EF}_r + \text{pPD1}$ was performed and yielded similar results (Extended Data Fig. 6). $\text{EF}_s + \text{pPD1}::\Delta\text{bacAB}$ is outcompeted by EF_r in all circumstances (Fig. 2b, e, h), confirming its loss of colonization fitness. pPD1 also provided a competitive advantage during *in vitro* co-culturing competitions (Extended Data Fig. 7). Complementation of the ΔbacAB strain restored bacteriocin activity, stable gut colonization (Extended Data Fig. 8) and the competitive advantage over other enterococci (Fig. 2c, f, i and Extended Data Fig. 8).

During competition experiments, we noted the persistence of EF_r when given in combination with $\text{EF}_s + \text{pPD1}$. This suggested that $\text{EF}_s + \text{pPD1}$ was able either to clear a niche and enhance overall EF_r survival or to transfer pPD1 to EF_r by conjugation. To determine whether conjugation of pPD1 occurred in the gastrointestinal tract we screened drinking water, gastrointestinal tissue (Supplementary Table 3)

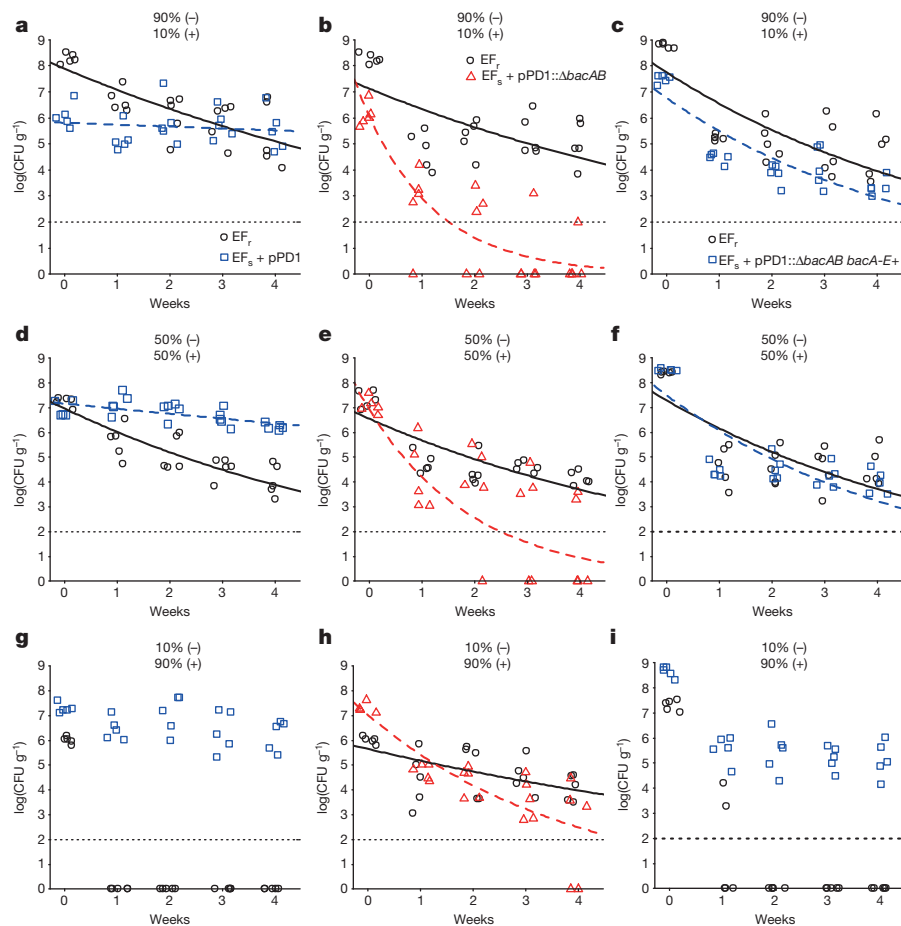


Figure 2 | Bacteriocin provides a competitive survival advantage to *E. faecalis* in the gastrointestinal tract. Mice ($n = 5$ per group) were given mixtures of EF_x (-) and $EF_x + pPD1$ (+) (a, d, g), EF_x and $EF_x + pPD1::\Delta bacAB$ (+) (b, e, h) or EF_x and $EF_x + pPD1::\Delta bacAB bacA-E$ (+) (c, f, i) in drinking water at the indicated ratios. Faecal shedding was determined weekly. Each symbol represents an individual animal. Lines are fitted using an exponential decay model. Rate of decay is significantly different between the two groups in all cases except c and f: a, $P = 0.0004$; b, $P < 0.0001$; d, $P < 0.0001$; e, $P = 0.001$; g, $P = 0.0117$ (week 0) and $P = 0.0075$ (weeks 1–4); h, $P = 0.008$; i, $P = 0.012$ (weeks 0 and 1) and $P = 0.0075$ (weeks 2–4). The results in a, d and g are representative of three biologically independent experiments; data in b, c, e, f, h and i are the results of one experiment. Black dotted lines (all panels) indicate the limit of detection at 100 CFU per g faeces.

and faecal isolates (Fig. 3a, b) of EF_x by colony polymerase chain reaction (PCR). No evidence of conjugation was observed in drinking water. The detection of transconjugants in the intestine or faeces indicated that conjugation occurred in the gastrointestinal tract. At week 1, approximately 25% of the EF_x faecal colonies selected had obtained the pPD1 plasmid via conjugation (Fig. 3a, b). By week 4, nearly 100% of the EF_x colonies harboured pPD1 (Fig. 3a, b) and produced bacteriocin (not shown). Similar results were observed in each segment of the lower gastrointestinal tract (Supplementary Table 3). $EF_x + pPD1::\Delta bacAB$ showed no evidence of the ability to conjugate its defective plasmid (not shown). Although complementation of *bacA-E* ectopically (pSK29) restored bac-21 production in $EF_x + pPD1::\Delta bacAB$, plasmid conjugation was not restored *in vitro* or *in vivo*. The mechanism underlying this observation is currently unknown.

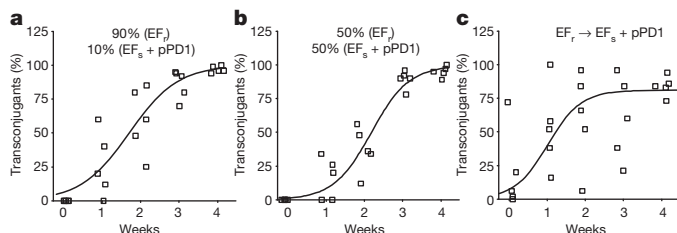


Figure 3 | pPD1 is transferred via conjugation within the mouse gastrointestinal tract. a, b, Mice ($n = 5$ per group) were colonized with mixtures of EF_x and $EF_x + pPD1$ in drinking water at the indicated ratios. Faecal EF_x colonies (100 per animal) were screened weekly for pPD1. c, Mice ($n = 5$ per group) were stably colonized with EF_x and then challenged with $EF_x + pPD1$. Faecal EF_x colonies (50 per animal) were screened weekly for pPD1. In a–c, each symbol represents the percentage of faecal transconjugants in an individual animal. The fitted logistic curve is overlain. The results in a–c are from one experiment each.

To determine whether $EF_x + pPD1$ could displace an established resident enterococcal population, we performed sequential colonization experiments, in which EF_x colonization was stably established, followed by challenge with $EF_x + pPD1$. The levels of EF_x were not notably altered after introducing $EF_x + pPD1$ (Extended Data Fig. 9); however, transconjugants of EF_x containing pPD1 started to appear as early as week 0 and neared 100% by week 4 (Fig. 3c). Although these results cannot distinguish extensive transfer of pPD1 by conjugation from a small number of transfer events followed by proliferation of pPD1-containing transconjugants, we conclude that pPD1 enhances *E. faecalis* competition in the gastrointestinal tract by bac-21-mediated killing or through transfer of a functional pPD1 plasmid via conjugation. To probe conjugation dynamics between non-isogenic enterococci, we investigated indigenous enterococcal transconjugants in mice that were colonized with $EF_x + pPD1$ (Fig. 1e and Extended Data Fig. 3b–e). We found that conjugation of pPD1 to indigenous enterococci is possible and can be observed both *in vivo* and *in vitro*, but also that it happens at a low and variable frequency (Extended Data Fig. 10). The reasons for this remain unknown, but are likely to be multifactorial, especially in the gastrointestinal tract.

Colonization with antibiotic-resistant enterococci precedes emergence of enterococcal infections. Bac-21-producing *E. faecalis* exhibits bacteriocin activity *in vitro* against many multidrug-resistant clinical isolates of *E. faecium* and *E. faecalis* (Supplementary Table 4). To determine whether bac-21 can disrupt *in vivo* colonization by vancomycin-resistant *E. faecalis* V583, mice colonized with rifampicin-resistant V583 ($V583_r$) were challenged with the complemented bac-21 + strain ($EF_x + pPD1::\Delta bacAB bacA-E$). The conjugation defect of this strain eliminated the risk for bacteriocin transfer to $V583_r$. $EF_x + pPD1::\Delta bacAB bacA-E$ successfully eliminated $V583_r$ levels to below the detection limit in most of the mice compared to

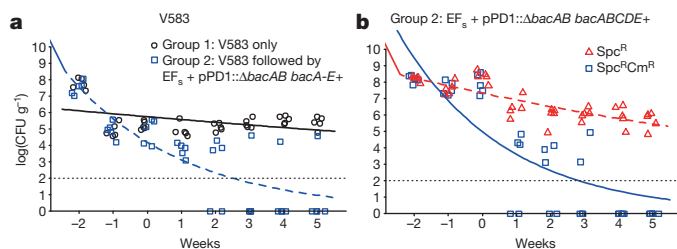


Figure 4 | Bacteriocin reduces V583_r colonization. Mice ($n = 5$ per group) were colonized with V583_r. V583_r was removed from the drinking water of both groups two weeks before sampling (week -2). Group 1 received sterile water, whereas group 2 received EF_s + pPD1::Δ*bacAB bacA-E* + in their drinking water for two additional weeks, followed by sterile water at week 0. Faecal levels of V583_r (a) and EF_s + pPD1::Δ*bacAB bacA-E* + (b) were enumerated weekly. The retention of complementation plasmid pAM401A::*bacA-E* + by EF_s + pPD1::Δ*bacAB bacA-E* + was determined weekly (b). Lines are fitted using an exponential decay model and the rate of decay is significantly different between the two groups in both a ($P < 0.0001$) and b ($P < 0.0001$). Each symbol represents an individual animal and data are representative of three biologically independent experiments. The black dotted lines indicate the limit of detection at 100 CFU per g faeces. Spc^R, spectinomycin resistant; Spc^RCm^R, spectinomycin resistant, chloramphenicol resistant.

the control group, which showed steady levels of V583_r in faeces (Fig. 4a). As expected, EF_s + pPD1::Δ*bacAB bacA-E* + did not transfer bacteriocin to V583_r during gut colonization and the loss of pAM401A::*bacA-E* resulted in persistent colonization by the bacteriocin-defective Δ*bacAB* strain (Fig. 4b). V583_r clones recovered from the one mouse with persistent V583_r levels in the treated group (group 2) were examined and found to be susceptible to bac-21 *in vitro*. We suggest that a stochastic, particularly rapid loss of complementing plasmid (pAM401A::*bacA-E*) from the challenging strain (EF + pPD1::Δ*bacAB bacA-E* +) in that particular mouse allowed V583_r to persist over time. These results demonstrate that introducing a conjugation-defective bac-21-producing strain into the gastrointestinal tract can successfully reduce colonization by multidrug-resistant *E. faecalis*.

Our findings provide proof of concept for a novel therapeutic strategy to specifically decolonize antibiotic-resistant enterococci from the gastrointestinal tract (of patients) and thereby prevent the emergence of resistant enterococcal infections that are otherwise difficult, or impossible, to treat. Although stable colonization by a bac-21-producing *E. faecalis* strain (EF + pPD1::Δ*bacAB bacA-E* +) can eliminate multidrug-resistant enterococci from the gut, we expect that the bacteriocin-producing strain may itself eventually acquire antibiotic resistance determinants, compromising its utility. Thus, therapeutic implementation of this strategy will probably require a modified approach. Nevertheless, use of a commensal bacterium to deliver bac-21 leverages the niche specificity of the therapeutic strain to produce bac-21 directly in the appropriate niche(s) at levels sufficient to inhibit other *E. faecalis*, yet not perturb the overall community. We expect that the concept presented here—specific inhibition of particular members of the gut ecosystem—will be widely applicable to modulate gut colonization by other problematic organisms as well.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 20 February; accepted 25 August 2015.

Published online 19 October 2015.

- Richards, M. J., Edwards, J. R., Culver, D. H. & Gaynes, R. P. Nosocomial infections in combined medical-surgical intensive care units in the United States. *Infect. Control Hosp. Epidemiol.* **21**, 510–515 (2000).

- Broaders, E., Gahan, C. G. & Marchesi, J. R. Mobile genetic elements of the human gastrointestinal tract: potential for spread of antibiotic resistance genes. *Gut Microbes* **4**, 271–280 (2013).
- Nes, I. F., Diep, D. B. & Ike, Y. in *Enterococci: From Commensals to Leading Causes of Drug-Resistant Infection* (eds Gilmore, M. S. et al.) Ch. 13 (Massachusetts Eye and Ear Infirmary, 2014).
- Fujimoto, S., Tomita, H., Wakamatsu, E., Tanimoto, K. & Ike, Y. Physical mapping of the conjugative bacteriocin plasmid pPD1 of *Enterococcus faecalis* and identification of the determinant related to the pheromone response. *J. Bacteriol.* **177**, 5574–5581 (1995).
- Hooper, L. V., Midtvedt, T. & Gordon, J. I. How host–microbial interactions shape the nutrient environment of the mammalian intestine. *Annu. Rev. Nutr.* **22**, 283–307 (2002).
- Turnbaugh, P. J. et al. The human microbiome project. *Nature* **449**, 804–810 (2007).
- Round, J. L. & Mazmanian, S. K. The gut microbiota shapes intestinal immune responses during health and disease. *Nature Rev. Immunol.* **9**, 313–323 (2009).
- Hidron, A. I. et al. NHSN annual update: antimicrobial-resistant pathogens associated with healthcare-associated infections: annual summary of data reported to the National Healthcare Safety Network at the Centers for Disease Control and Prevention, 2006–2007. *Infect. Control Hosp. Epidemiol.* **29**, 996–1011 (2008).
- Shepard, B. D. & Gilmore, M. S. Antibiotic-resistant enterococci: the mechanisms and dynamics of drug introduction and resistance. *Microbes Infect.* **4**, 215–224 (2002).
- Paulsen, I. T. et al. Role of mobile DNA in the evolution of vancomycin-resistant *Enterococcus faecalis*. *Science* **299**, 2071–2074 (2003).
- Brandl, K. et al. Vancomycin-resistant enterococci exploit antibiotic-induced innate immune deficits. *Nature* **455**, 804–807 (2008).
- Ubeda, C. et al. Vancomycin-resistant *Enterococcus* domination of intestinal microbiota is enabled by antibiotic treatment in mice and precedes bloodstream invasion in humans. *J. Clin. Invest.* **120**, 4332–4341 (2010).
- Dobson, A., Cotter, P. D., Ross, R. P. & Hill, C. Bacteriocin production: a probiotic trait? *Appl. Environ. Microbiol.* **78**, 1–6 (2012).
- Hibbing, M. E., Fuqua, C., Parsek, M. R. & Peterson, S. B. Bacterial competition: surviving and thriving in the microbial jungle. *Nature Rev. Microbiol.* **8**, 15–25 (2010).
- Corr, S. C. et al. Bacteriocin production as a mechanism for the anti-infective activity of *Lactobacillus salivarius* UCC118. *Proc. Natl Acad. Sci. USA* **104**, 7617–7621 (2007).
- Perez, R. H., Zendo, T. & Sonomoto, K. Novel bacteriocins from lactic acid bacteria (LAB): various structures and applications. *Microb. Cell Fact.* **13** (Suppl 1), S3 (2014).
- Grande Burgos, M. J., Pulido, R. P., Del Carmen Lopez Aguayo, M., Galvez, A. & Lucas, R. The cyclic antibacterial peptide enterocin AS-48: isolation, mode of action, and possible food applications. *Int. J. Mol. Sci.* **15**, 22706–22727 (2014).
- Cruz, V. L., Ramos, J., Melo, M. N. & Martinez-Salazar, J. Bacteriocin AS-48 binding to model membranes and pore formation as revealed by coarse-grained simulations. *Biochim. Biophys. Acta* **1828**, 2524–2531 (2013).
- Tomita, H., Fujimoto, S., Tanimoto, K. & Ike, Y. Cloning and genetic and sequence analyses of the bacteriocin 21 determinant encoded on the *Enterococcus faecalis* pheromone-responsive conjugative plasmid pPD1. *J. Bacteriol.* **179**, 7843–7855 (1997).
- Gálvez, A., Maqueda, M., Martínez-Bueno, M. & Valdivia, E. Bactericidal and bacteriolytic action of peptide antibiotic AS-48 against Gram-positive and Gram-negative bacteria and other organisms. *Res. Microbiol.* **140**, 57–68 (1989).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank I. Banla for constructing the IB1 (V583_r) strain. We are grateful to J. Barbieri and the members of the Salzmann and Kristich laboratories for critical review of the manuscript. We thank M.S. Gilmore for providing enterococcal strains. This work was supported by grants from the National Institutes of Health: AI057757 (N.H.S.), AI097619 (N.H.S.), GM099526 (N.H.S.), AI081692 (C.J.K.) and OD006447 (C.J.K.).

Author Contributions S.K., C.J.K. and N.H.S. designed and conceived the study. S.K. performed most of the experiments and the analysis; D.J.B. and M.H. assisted in the development of the colonization model. V.L. performed bioinformatics analysis. R.C. and P.B. contributed to the sequential colonization experiment. P.S. and Y.C. performed statistical analysis; S.K., C.J.K. and N.H.S. interpreted the data and wrote the manuscript. C.J.K. and N.H.S. secured funding.

Author Information The complete sequence of pPD1 was deposited in GenBank under accession number KT290268. 16S rDNA sequences generated for microbiome analyses are deposited in the NCBI-SRA archive under study accession number SRP061808 and BioProject accession number PRJNA290480. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to N.H.S. (nsalzman@mcw.edu) and C.J.K. (ckristich@mcw.edu).

METHODS

Bacterial strains, growth media and chemicals. The strains used in this study are listed in Supplementary Table 5. Brain-heart infusion medium (BHI) and m-Enterococcus agar (Ent-agar) (Difco) were prepared as described by the manufacturer (Becton Dickinson). Antibiotics were purchased from Sigma and used at the following concentrations: rifampicin, 200 µg ml⁻¹; spectinomycin, 500 µg ml⁻¹; erythromycin, 10 µg ml⁻¹; chloramphenicol, 15 µg ml⁻¹; tetracycline, 10 µg ml⁻¹. *E. faecalis* was cultured in BHI media at 37 °C. All restriction enzymes were purchased from New England Biolabs. Phusion High-Fidelity DNA Polymerase (Thermo Scientific) was used for all PCRs performed for strain and plasmid construction. Oligonucleotides were synthesized by Fisher Scientific (Supplementary Table 6).

Animals. The committee for animal care and use at the Medical College of Wisconsin approved all animal-related procedures and experiments. Five-week-old male FVB and C57BL/6 mice were obtained from Taconic Laboratories (Oxnard and Germantown facilities). Upon arrival, mice were allowed to adapt to the new environment for at least 1 week before the start of any experiment. Animals were housed under specific-pathogen-free conditions in the Medical College of Wisconsin vivarium. Experimental sample sizes were determined by appropriate husbandry considerations as determined by the Medical College of Wisconsin vivarium and experiments were repeated as described. No blinding was performed and no scheme of randomization was applied when allocating mice for the experiments.

Generation of *bacAB* deletion in pPD1. A mutant lacking *bacAB* was constructed using a marker gene (*ermC*) insertion in an in-frame deletion. The design for in-frame deletion and the procedure for mutant construction were similar to those used previously with pCJ218 (ref. 21). The PCR products with the first and last codons in the *bacAB* fragment and flanking the *ermC* gene were inserted between the XbaI and SphI sites of pCJ218 by Gibson Assembly²² to generate pSK21. Deletion mutants were isolated by plating on counter-selection medium containing 4-chloro-DL-phenylalanine followed by prolonged incubation at 30 °C (ref. 21). The desired mutants harbouring the *bacAB* deletion in pPD1 plasmids (pSK35) were identified by PCR analysis and erythromycin selection. Two independent isolates were obtained and were analysed to exhibit identical phenotypes.

Complementation studies. The *bacABCDE* fragment was amplified using primers listed in Supplementary Table 6 and was cloned into a pAM401 vector using the Gibson assembly²² approach, resulting in pSK29. Strains with a *bacAB* deletion in pPD1 (CK135 pSK35 or OG1sp pSK35) were electroporated with pSK29. PCR analysis, chloramphenicol resistance and bacteriocin assay confirmed the complementation construct.

Bacteriocin assays. As previously described¹⁹, 50 µl of an overnight culture of the indicator strain (CK135 or OG1sp) grown in BHI broth was added to 5 ml of molten BHI soft agar and then spread evenly onto a BHI agar plate. After solidification, 2 µl of the test strain that was grown overnight was spotted onto the soft agar. Zones of inhibition of the susceptible strain around the spots were monitored after overnight incubation at 37 °C.

Mouse colonization model. Overnight cultures of *E. faecalis* were washed with sterile water and added to autoclaved water to a final concentration of 5 × 10⁸ CFU ml⁻¹. Persistence of *E. faecalis* in drinking water was determined daily and remained between 10⁷ and 10⁸ CFU ml⁻¹ over 3 days. Drinking water was changed every 3–4 days to maintain the appropriate inoculum and mice were allowed to drink *ad libitum*. After two weeks, the inoculated drinking water was replaced with sterile water for the duration of the experiment. For the sequential colonization experiment, mice were first colonized with a strain lacking pPD1 (initial strain) by feeding the bacteria through drinking water for two weeks. Three days after the initial strain was withdrawn from drinking water, the challenge strain was introduced in the drinking water for two weeks, after which animals were returned to regular sterile water (week 0).

Bacterial culture and quantification of *E. faecalis* from mouse faeces and intestines. Faecal pellets obtained directly from living mice were weighed and homogenized in 1 ml PBS. Faecal homogenates were plated in serial dilution on BHI agar plates with appropriate antibiotics to enumerate levels of *E. faecalis* colonization. After mice were euthanized, the intestinal tract was removed and divided into three segments: distal 10 cm of the small intestine, caecum and large intestine. Each segment was homogenized in 2 ml PBS and cultured on BHI media with appropriate antibiotics. To enumerate indigenous enterococci or total enterococci, faecal and tissue homogenates were plated in serial dilution on Ent-agar (Difco). Alternative to BHI, laboratory strains were enumerated on Ent-agar supplemented with appropriate antibiotics.

Colony PCR for *bacA* and *bacD* genes. To determine transconjugants, colonies were selected and used as PCR templates for identification of the *bacA* or *bacD*

gene using gene-specific primers as described in Supplementary Table 6. PCR products were identified using agarose gel electrophoresis.

Plasmid DNA preparation and sequencing of pPD1. The pPD1 plasmid was purified as described previously²³. Sequencing was performed at GENEWIZ, Inc. (South Plainfield, New Jersey) on the Illumina HiSeq Platform in a 2 × 150-bp paired-end configuration. The *de novo* assembly using the CLC Genomics Server 6.5.1 was used to obtain assembled contigs/scaffolds. In total, 59 ORFs larger than 100 bp showing sequence similarity to known sequences in the NCBI database were identified by the NCBI ORF finder (<http://www.ncbi.nlm.nih.gov/projects/gorf/>).

Bacterial genomic DNA extraction. Caeca were isolated from experimental animals and homogenized as described²⁴. Genomic DNA was extracted from tissues using the MO BIO PowerFecal DNA Isolation kit (MO BIO, Carlsbad, California) with slight modification in the protocol. After addition of solution C1 and heating the samples at 65 °C for 10 min, the sample was subjected to further heating at 95 °C for 10 min followed by vigorous bead beating using PowerLyzer (MO BIO).

Amplification of bacterial 16S gene sequences and high-throughput sequencing. The 16S rDNA V4 region amplicons (single index) were produced by PCR and sequenced on the MiSeq platform (Illumina) using the 2 × 250-bp protocol yielding paired-end reads that overlap by ~247 bp (ref. 25) (performed by Metanome, Baylor College of Medicine, Waco, Texas). Following sequencing, raw BCL (binary base call) files were retrieved from the MiSeq platform and called into fastq files by Casava v1.8.3 (Illumina). The read pairs were demultiplexed based on unique molecular barcodes, filtered for PhiX using Bowtie2 v2.2.1 (ref. 26) and reconstituted into two fastq files for each read using standard BASH. Sequencing reads were merged (allowing 4 mismatches per 50 bases or more) and processed using USEARCH v7.0.1001 (ref. 27). Sequences were demultiplexed using QIIME v1.8.0 (ref. 28) and then clustered using the UPARSE pipeline²⁷. Operational taxonomic unit (OTU) classification was achieved by mapping the UPARSE OTU table to the SILVA database²⁹. Abundances were recovered by mapping the demultiplexed reads to the UPARSE OTUs. A custom script constructed an OTU table from the output files generated in the two previous steps. The OTU table was used to calculate alpha and beta diversities and provide taxonomic summaries^{30–32}.

Bioinformatic analysis. Microbiome data were analysed using the Vegan³³ and Ecodist³⁴ packages in R 3.0.2 (ref. 35). Sequence counts of each sample were normalized to the average sequencing depth, and the Bray–Curtis metric was used to assess intersample (beta) diversity. Statistical significance for differences in microbiome diversity between groups was determined using Adonis (in Vegan). NMDS ordination (in Ecodist) was used to visualize group clustering and diversity distance between samples. Log-transformed abundance data (log₁₀ of sequence count + 1) was used to normalize the data and statistically significant differences in OTU abundance between groups were determined by heteroscedastic, two-sided Student's *t*-tests.

Statistical analysis of data. To compare the *E. faecalis* in faeces/tissues under different experimental conditions, we applied either a non-parametric Mann–Whitney–Wilcoxon or Kruskal–Wallis test owing to the skewness of the data. To model the response changes over time, we fitted an exponential decay in the following form:

$$EF \text{ in faeces (log CFU per g)} = k_0 e^{-k_1 t}$$

where k_0 is the *E. faecalis* in faeces (log CFU per g) at time 0, k_1 is the rate of decay and t is time.

We also fitted a logistic curve to model the percentage transconjugants using the following equation:

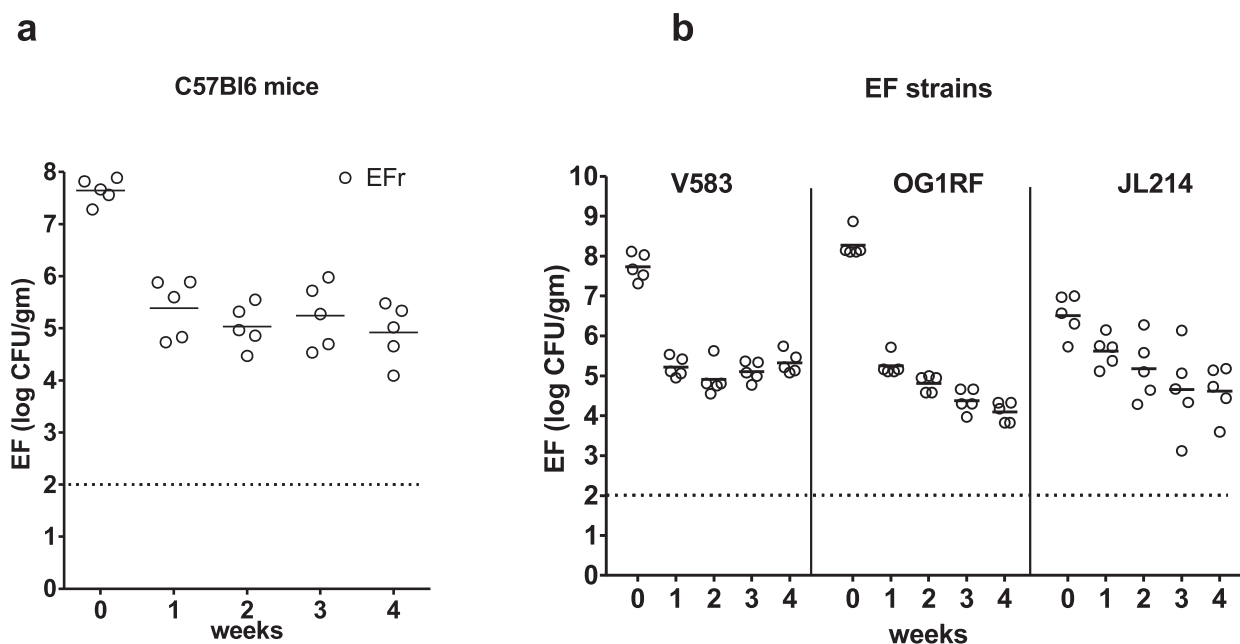
$$\text{Percentage transconjugants (\%)} = \frac{k_1}{1 + e^{-k_2(t - k_3)}}$$

where k_1 is the maximum value of the curve, k_2 is the steepness of the curve, k_3 is the time of the midpoint of the sigmoid and t is time.

For the non-linear mixed-effects models, we used a compound symmetry correlation structure to model the dependency among observations. If compound symmetry was not a good fit, a general correlation matrix or AR(1) (autocorrelation structure of order 1) correlation matrix was used. For responses that had non-zero values only at one or two time points, a model could not be fit. In these cases, we compared the responses at each time point. A linear mixed-effect model with autoregressive covariance structure was used to compare the *E. faecalis* in faeces over time for *E. faecalis* and *E. faecalis* + pPD1. Mice were treated as random. The response was log transformed and groups were compared using a two-tailed *t*-test at an alpha of 0.05. With 5 mice in each group, we had at least 80% power to detect a difference of at least 2.4 standard deviations. PASS 2008 was used for the calculation (<http://www.ncss.com>). SAS (SAS Institute, Cary, North Carolina),

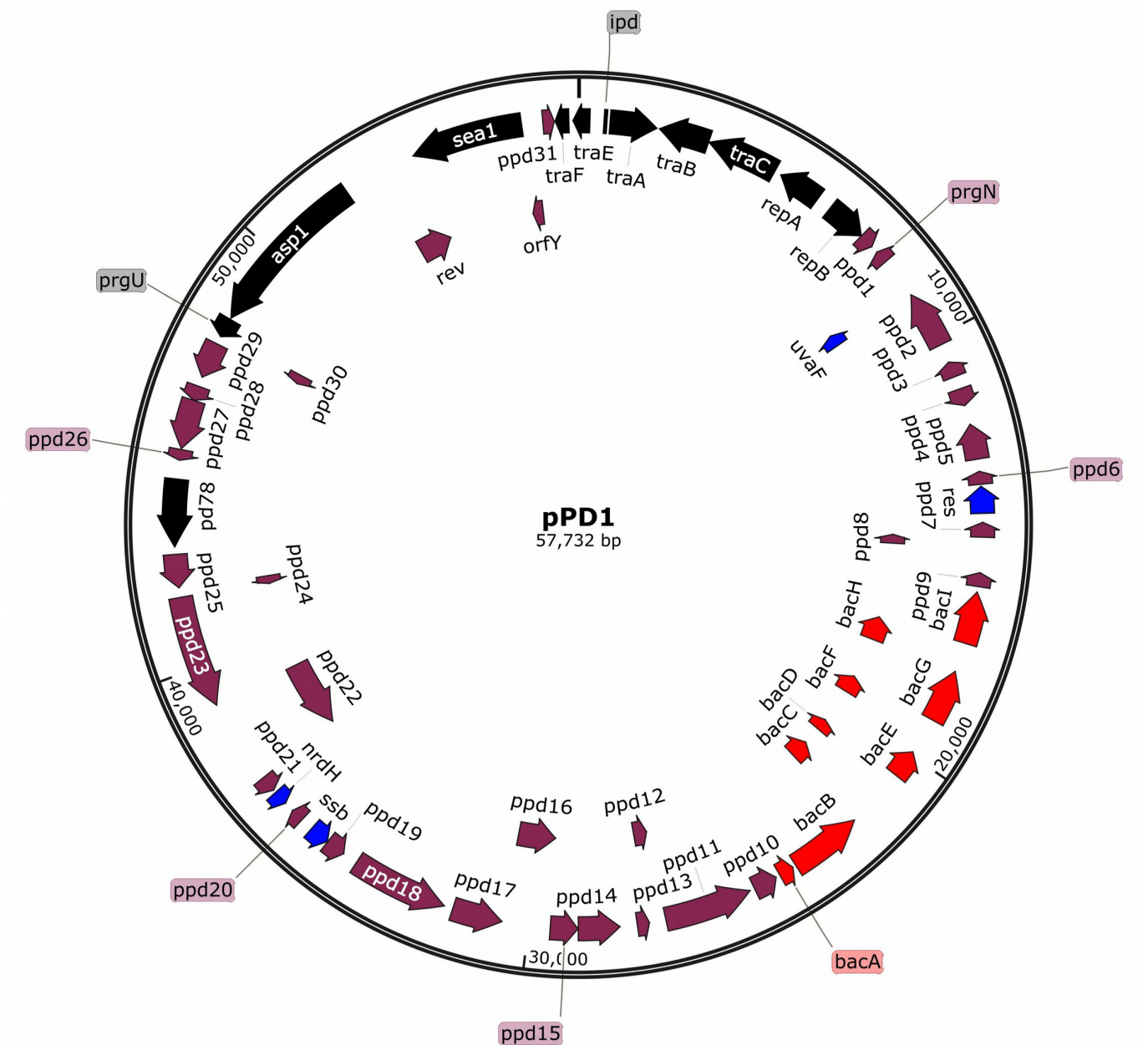
S-PLUS (Insightful Corporation, Seattle, Washington), R (ref. 35) and SPSS (IBM Corporation, Armonk, New York) were used for statistical analyses. All tests were two-tailed. A *P*-value <0.05 was considered statistically significant.

21. Vesić, D. & Kristich, C. J. A Rex family transcriptional repressor influences H₂O₂ accumulation by *Enterococcus faecalis*. *J. Bacteriol.* **195**, 1815–1824 (2013).
22. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature Methods* **6**, 343–345 (2009).
23. O'Sullivan, D. J. & Klaenhammer, T. R. Rapid mini-prep isolation of high-quality plasmid DNA from *Lactococcus* and *Lactobacillus* spp. *Appl. Environ. Microbiol.* **59**, 2730–2733 (1993).
24. Croswell, A., Amir, E., Tegatz, P., Barman, M. & Salzman, N. H. Prolonged impact of antibiotics on intestinal microbial ecology and susceptibility to enteric *Salmonella* infection. *Infect. Immun.* **77**, 2741–2753 (2009).
25. Caporaso, J. G. *et al.* Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.* **6**, 1621–1624 (2012).
26. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357–359 (2012).
27. Edgar, R. C. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods* **10**, 996–998 (2013).
28. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7**, 335–336 (2010).
29. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2013).
30. Lozupone, C., Hamady, M. & Knight, R. UniFrac – an online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinformatics* **7**, 371 (2006).
31. Chao, A., Chazdon, R. L., Colwell, R. K. & Shen, T. J. Abundance-based similarity indices and their estimation when there are unseen species in samples. *Biometrics* **62**, 361–371 (2006).
32. Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).
33. Oksanen, J. *et al.* Vegan: Community Ecology Package. R package version 2.0-10. <https://cran.r-project.org/web/packages/vegan/index.html> (2013).
34. Goslee, S. C. & Urban, D. L. The ecodist package for dissimilarity-based analysis of ecological data. *J. Stat. Softw.* **22**, 1–19 (2007).
35. R. Development Core Team. R: a language and environment for statistical computing. <http://www.R-project.org/> (R Foundation for Statistical Computing, 2014).



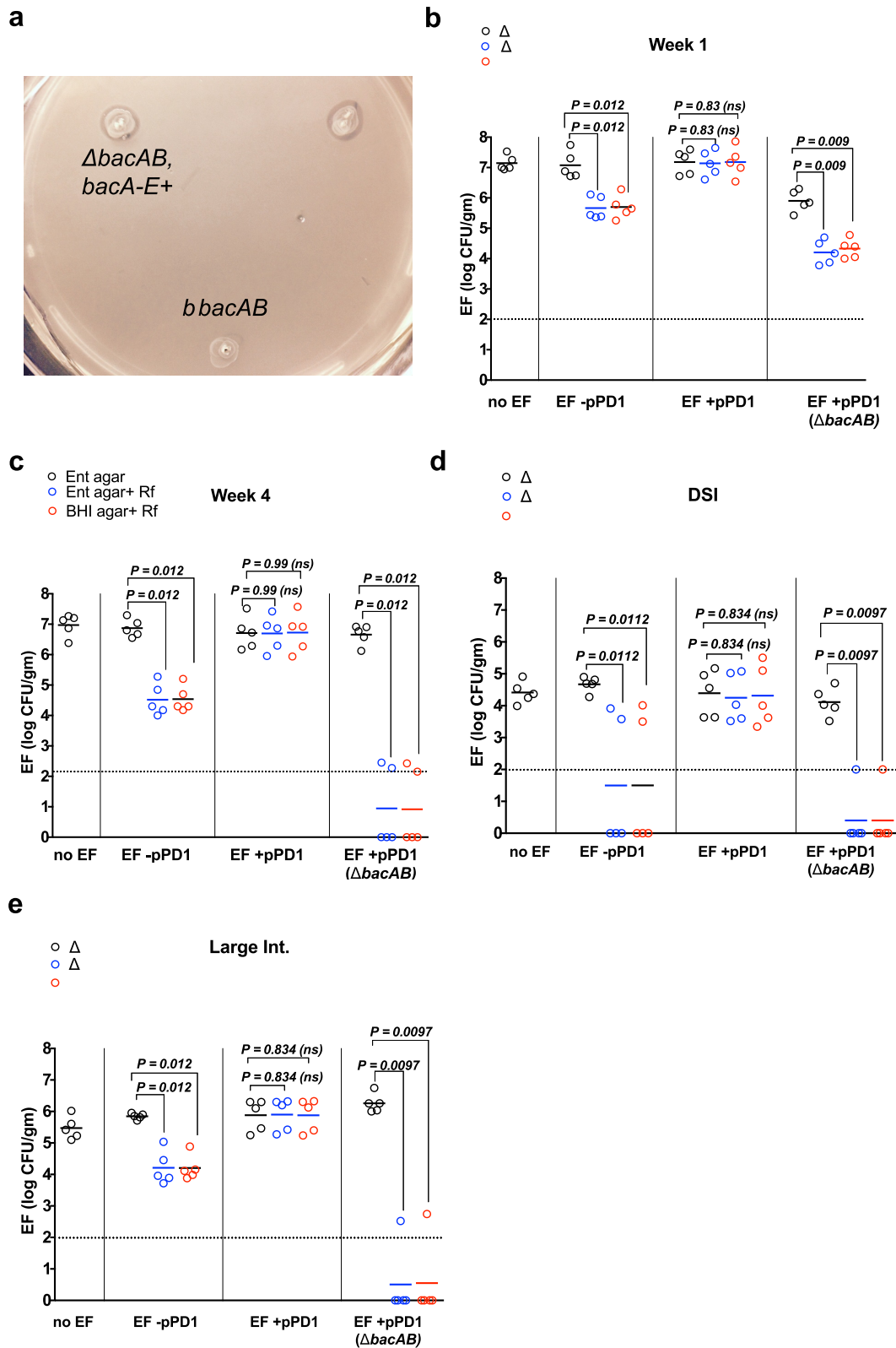
Extended Data Figure 1 | EF colonization. **a**, C57Bl/6 mice ($n = 5$) were given EFr in drinking water for 14 days. Faecal samples were taken from each animal at the transition to sterile drinking water (week 0) and then weekly. EFr abundance was determined by enumeration on brain-heart infusion (BHI) agar with rifampicin. **b**, V583, OG1RF and JL214 strains of *E. faecalis* that are rifampicin resistant were fed to groups of mice ($n = 5$ per group). *E. faecalis*

strains were enumerated weekly as described above. In both **a** and **b**, horizontal lines indicate geometric means and each symbol represents an individual animal. Data are representative of more than three experiments in **a**; in **b**, data are representative of more than three experiments for V583_r and OG1RF and the result of one experiment for JL214.



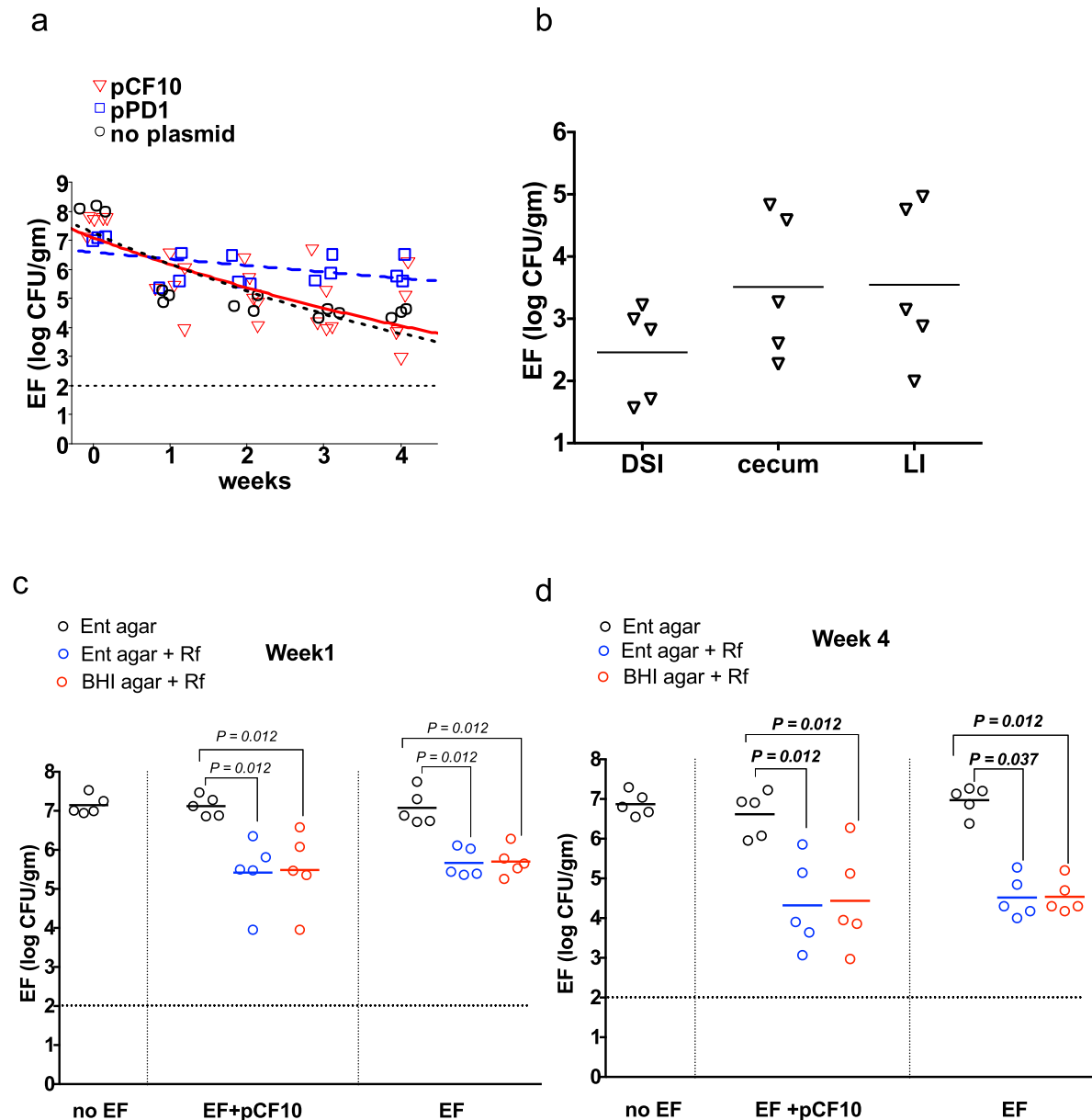
Extended Data Figure 2 | Map of plasmid pPD1. The 59 ORFs identified in the nucleotide sequence of pPD1 are located on a circular map. Arrows indicating the direction of transcription show the ORFs. Different colours indicate coding regions for conjugation (black), *bac* operon (red) and maintenance or repair (blue). Hypothetical coding regions are shown in

magenta. A circular plasmid map was generated using the SnapGene software (GSL Biotech; available at <http://www.snapgene.com>). Schematic diagrams of multiple alignments of plasmids were produced by manually realigning the linear plasmid maps generated by the SnapGene Viewer.



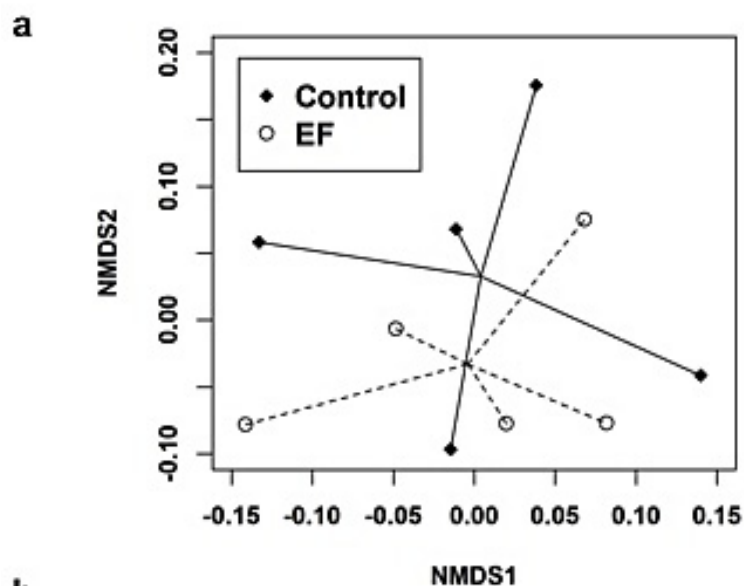
Extended Data Figure 3 | pPD1 enhances *E. faecalis* competition for an intestinal niche. **a**, Bacteriocin assay by the soft agar method, with EF_r + pPD1:: $\Delta bacAB$ $bacA-E+$, EF_r + pPD1 and EF_r + pPD1:: $\Delta bacAB$ spotted on a lawn of susceptible *E. faecalis*. Mice ($n = 5$ per group) were given EF_r, EF_r + pPD1, EF_r + pPD1:: $\Delta bacAB$ or sterile drinking water for 14 days, after which all mice were given sterile water. **b**, **c**, One week (**b**) and four weeks (**c**) after withdrawal of *E. faecalis* from the drinking water, faecal samples were collected and the abundance of total enterococci was determined by

enumeration on m-Enterococcus selective agar (Ent agar). Laboratory strains of *E. faecalis* were enumerated using Ent agar with rifampicin (Rf) or BHI agar with rifampicin. **d**, **e**, At the end of week 4, animals were euthanized and abundance of *E. faecalis* was determined in the distal small intestine (**d**) and large intestine (**e**) (also see Fig. 1e). The results shown are representative of two biologically independent experiments. In **b**–**e**, horizontal lines indicate geometric means and each symbol represents an individual animal.



Extended Data Figure 4 | EF + pPD1 but not EF_r + pCF10 dominates the intestinal enterococcal population. pCF10 is a well-studied pheromone-inducible conjugative plasmid of *E. faecalis* that encodes resistance to tetracycline but does not encode a known bacteriocin determinant. **a**, EF_r ($n = 3$ mice, no plasmid), EF_r + pPD1 ($n = 3$ mice, pPD1) or EF_r + pCF10, ($n = 5$ mice, pCF10) was added to the drinking water for 14 days and then replaced by sterile drinking water. Faecal samples were taken from each animal at the transition to sterile drinking water (week 0) and then weekly. *E. faecalis* abundance was determined by enumeration on BHI agar with rifampicin. An exponential decay model is used to fit the data and there are significant differences between the groups pCF10 (red) and pPD1 (blue) ($P = 0.012$), as well as between the pPD1-containing and plasmid-free groups (black) ($P = 0.007$). **b**, Four weeks after withdrawal of *E. faecalis* from the drinking

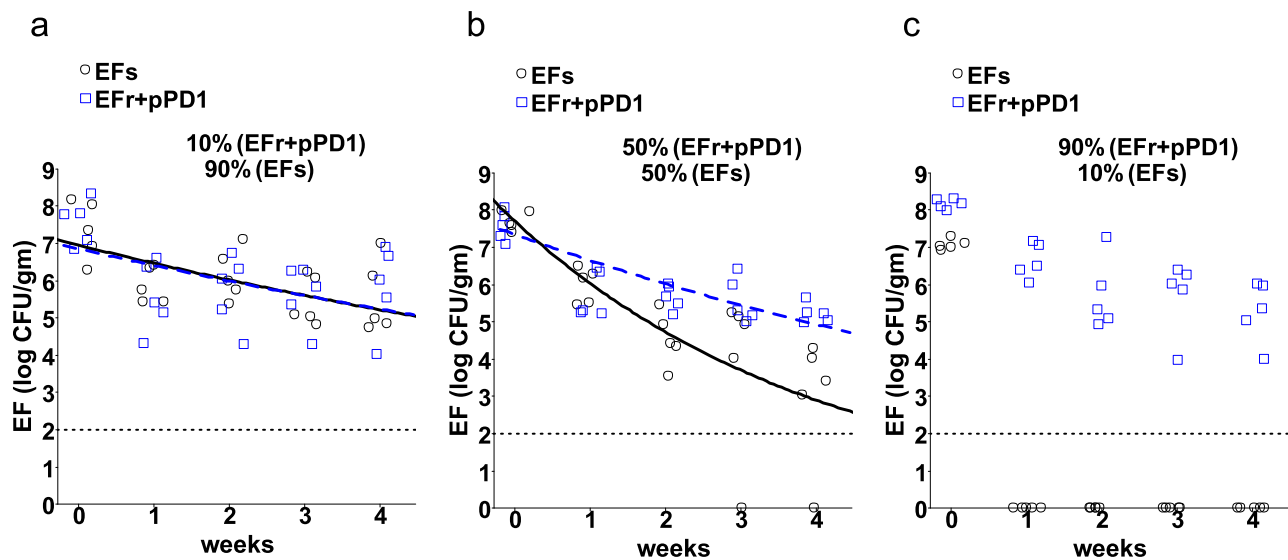
water, animals were euthanized and abundance of EF_r + pCF10 was determined in the distal small intestine (DSI), caecum and large intestine (LI). Mice colonized with EF_r + pCF10 maintained long-term faecal shedding of EF_r + pCF10 similar to EF_r and persistent colonization throughout the gastrointestinal tract. **c**, **d**, Abundance of enterococci in the faeces was determined by enumeration on m-Enterococcus agar (Ent agar), Ent agar with rifampicin or BHI agar with rifampicin at week 1 (**c**) and week 4 (**d**). Unlike EF_r + pPD1, which dominated the enterococcal niche in the gastrointestinal tract (Fig. 1e), EF_r + pCF10 did not outcompete the indigenous enterococci, colonizing at levels comparable to EF_r. In **b–d**, horizontal lines indicate geometric means and each symbol represents an individual animal; data were obtained from one experiment.



Taxon	EF vs EF +pPD1 t-test	Magnitude: (EF+pPD1) - (EF)
Family level		
Bacteria;__Deferribacteres;__Deferribacteres;__Deferribacterales;__Deferribacteraceae	0.00010	-1.04280
Bacteria;__Firmicutes;__Clostridia;__Clostridiales;__Deffluviitaleaceae	0.04195	-0.21928
Genus level		
Bacteria;__Deferribacteres;__Deferribacteres;__Deferribacterales;__Deferribacteraceae;__Mucispirillum;Other	0.00010	-1.04280
Bacteria;__Firmicutes;__Clostridia;__Clostridiales;__Lachnospiraceae;__Incertae_Sedis;Other	0.00545	-0.29577
Bacteria;__Firmicutes;__Clostridia;__Clostridiales;__Lachnospiraceae;__Blautia;Other	0.03635	0.14052
Bacteria;__Firmicutes;__Clostridia;__Clostridiales;__Lachnospiraceae;__Incertae_Sedis;Other	0.03877	-0.23034

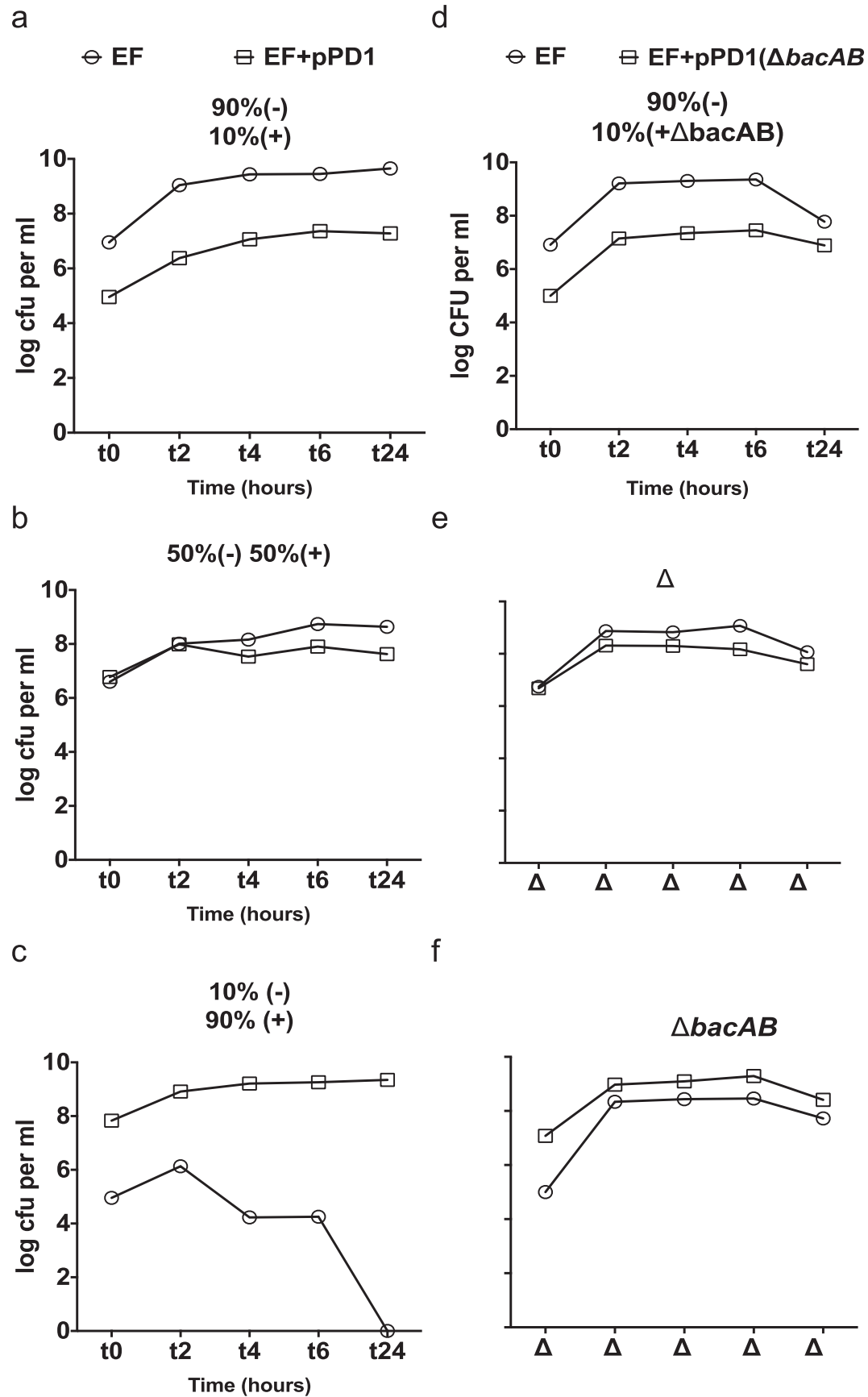
Extended Data Figure 5 | Microbiome analysis. **a**, NMDS ordination of control (no EF treatment) and *E. faecalis* samples as separated by the Bray-Curtis beta diversity metric. Control ($n = 5$ mice) and *E. faecalis* samples ($n = 5$ mice) are intermixed. No significant difference in beta diversity was seen between the two groups. Adonis P -value = 0.298. Samples are connected with lines to help visualize grouping. **b**, Analysis between *E. faecalis* and EF + pPD1 samples at the level of operational taxonomic unit suggests changes in the abundance of four bacterial genera, in particular for Deferribacteraceae;

Mucispirillum and Lachnospiraceae; *Incertae sedis* ($P = 0.0001$ and $P = 0.005$, respectively; heteroscedastic two-sided Student's t -test). Bacteria belonging to these two genera were tenfold and twofold lower, respectively, in EF + pPD1-colonized mice. (Note that the magnitude of change is shown in \log_{10} scale.) Changes in Deffluviitaleaceae; *Incertae sedis* and Lachnospiraceae; *Blautia* were not as pronounced ($P = 0.03$; heteroscedastic two-sided Student's t -test). However, analyses at the family taxonomy level suggest that the change in Deferribacteraceae was statistically significant.



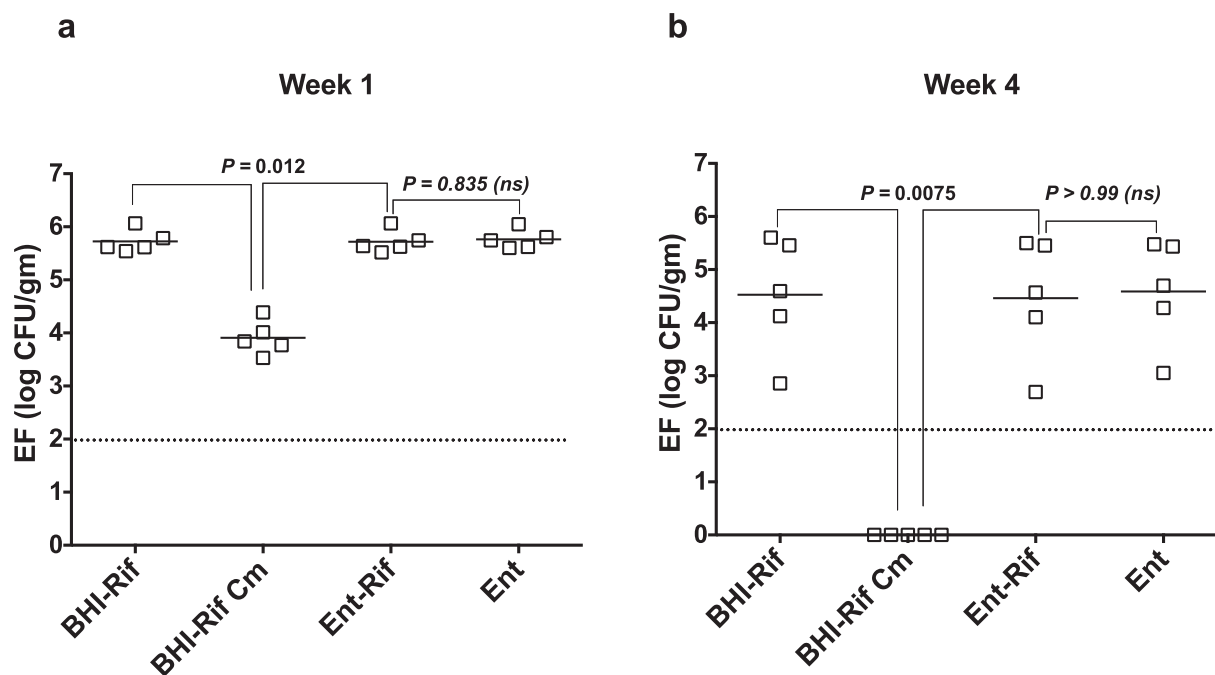
Extended Data Figure 6 | Reciprocal experiment for data from Fig. 2a, d and g. Groups of mice ($n = 5$ mice per group) were given mixtures of EF_r + pPD1 and EF_s in drinking water at ratios of 10%/90% (a), 50%/50% (b) and 90%/10% (c), respectively. Faecal samples were obtained at the transition to sterile drinking water (week 0) and then weekly. Abundance of each *E. faecalis* strain in the faeces was determined by enumeration on BHI agar with rifampicin and BHI agar with spectinomycin. Abundance of EF_r + pPD1 is

indicated by open squares; abundance of EF_s is indicated by open circles. Each symbol represents an individual animal. The differences between the two groups at each week were compared using a non-parametric Wilcoxon test. In a, no P -values were significant, in b, $P < 0.005$ and in c, $P = 0.0122$ (week 0) and $P = 0.0075$ (week 1–4). An exponential decay model is used for fitting the data in a and b. The results in a–c are from one experiment.



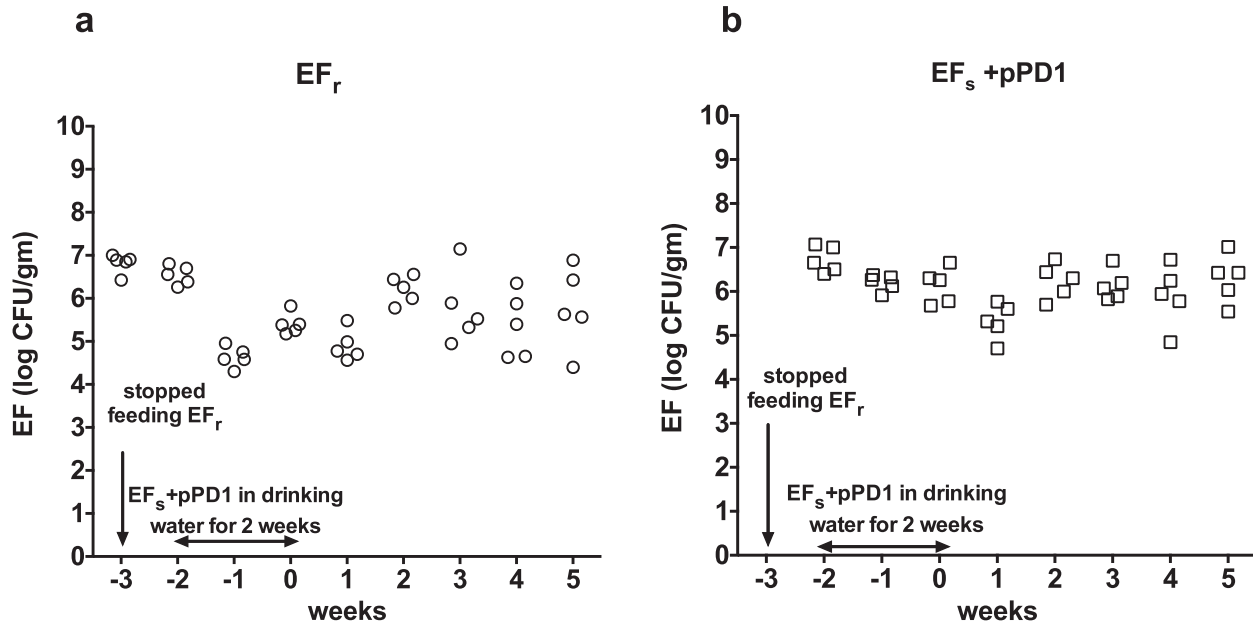
Extended Data Figure 7 | pPD1-associated competition *in vitro*. Three independent cultures were carried out with various mixed populations of EF_r (–) and EF_s + pPD1 (+) or EF_r (–) and EF_s + pPD1::Δ*bacAB* (+Δ*bacAB*) in 10 ml BHI broth at ratios of 90%/10% (**a**, **d**), 50%/50% (**b**, **e**) and 10%/90% (**c**, **f**). Samples for serial dilution were taken at 0, 2, 4, 6 and 24 h after the start of the experiment. Abundance of each *E. faecalis* strain in faeces was

determined by enumeration on BHI agar with rifampicin and BHI agar with spectinomycin. Evidence of conjugation was observed in *in vitro* co-cultures of **a** and **b** only by screening for transconjugants (EF_r + pPD1) via colony PCR. Open squares represent abundance of EF_s + pPD1 (**a–c**) or EF_s + pPD1::Δ*bacAB* (**d–f**). Open circles represent abundance of EF_r in all panels. Data are representative of two biologically independent experiments.



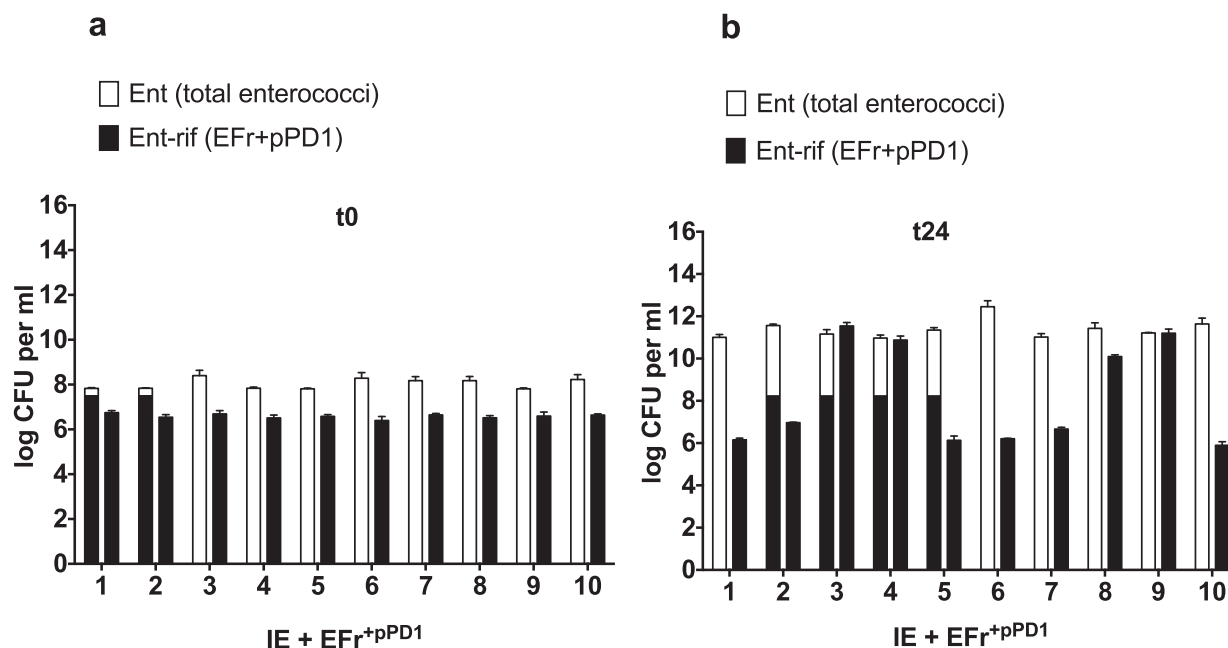
Extended Data Figure 8 | Complementation of bac-21 production restores colonization phenotype by providing a competitive advantage. Bacteriocin activity was restored upon ectopic expression of *bacA* to *bacE* (from pAM401) in EF + pPD1:: Δ *bacAB* but not in *E. faecalis* lacking pPD1, indicating that the distal part of the *bac* operon (*bacF* to *bacI*) is necessary for bacteriocin expression and that the *bacAB* in-frame deletion is not polar on downstream genes. Mice ($n = 5$) were given EF_s + pPD1:: Δ *bacAB* *bacABCDE* + as described in the methods and abundance was determined by enumeration on m-Enterococcus (Ent) agar, m-Enterococcus agar plus rifampicin (Ent-Rif) or BHI agar with rifampicin (BHI-Rif). The presence of pAM401A::*bacABCDE* + (complementing plasmid) was determined by enumerating CFU on BHI agar with rifampicin and chloramphenicol

(BHI-Rif Cm). Faecal samples were obtained at week 1 (a) and week 4 (b) after transition to sterile drinking water. Horizontal lines indicate geometric means. Each symbol represents an individual animal and data are representative of two biologically independent experiments. EF_s + pPD1:: Δ *bacAB* *bacABCDE* + stably colonized the gastrointestinal tract (a); however, in the absence of chloramphenicol selection, pAM401::*bacABCDE* was gradually lost from the population (b). Over time, loss of pAM401::*bacABCDE* resulted in the complemented strain reverting to the bacteriocin-defective Δ *bacAB* strain, with the loss of bacteriocin activity. Nevertheless, this strain persisted in the gut, suggesting that bac-21 was essential for clearing a niche for *E. faecalis*; once cleared, *E. faecalis* uses other mechanisms to maintain colonization.



Extended Data Figure 9 | EF_r levels are not altered by sequential colonization with $EF_s + pPD1$. Groups of mice ($n = 5$ mice per group) were given EF_r in drinking water for two weeks and subsequently challenged with $EF_s + pPD1$ in drinking water for another two weeks (starting from week -2)

before transition to sterile water (week 0). Faecal samples were obtained weekly to enumerate the abundance of EF_r (a) and $EF_s + pPD1$ (b) on BHI agar with rifampicin and spectinomycin, respectively. Each symbol represents an individual animal and data are from one experiment.



Extended Data Figure 10 | Conjugation frequency of pPD1 between the laboratory strain (EF + pPD1) and indigenous enterococci *in vitro*. To understand conjugation dynamics between non-isogenic species of enterococci, we investigated indigenous enterococcal transconjugants in mice that were colonized with EFr + pPD1 (Fig. 1e and Extended Data Fig. 3b–e). However, we were unable to detect rifampicin-sensitive enterococci from the faecal sample at week 4 (Extended Data Fig. 3c). At week 1, only nine clones of rifampicin-sensitive enterococci (out of 730 enterococci) were isolated from three mice in a group of five (Extended Data Fig. 3b). Bacteriocin assays and probing for the *bacA* gene sequence confirmed that six of the nine clones were transconjugants and were bac-21 positive. 16S rDNA gene sequencing of these nine clones showed notable similarities to *E. faecalis* 16S rDNA. To understand the reason for the low conjugation frequency between indigenous enterococci and the laboratory strain of EFr + pPD1, *in vitro* conjugation experiments were performed to assess the frequency of plasmid transfer. Ten new clones of indigenous enterococci were isolated from the faeces of five mice (two clones per mouse that were not colonized with any laboratory strain) by culturing on m-Enterococcus agar. EFr + pPD1 was mixed with each of

the ten indigenous enterococci clones in BHI broth at a ratio of 1:9. Samples for serial dilution were taken at the start of the culture (a) and 24 h after the start of the experiment (b). Abundance of total enterococci was determined using m-Enterococcus agar and BHI agar with rifampicin for EFr + pPD1. Data are representative of three biologically independent experiments. *In vitro* bacteriocin assays revealed that EFr + pPD1 is capable of killing most of the non-pPD1-containing indigenous enterococci strains (not shown). *In vitro* conjugation assays between individual indigenous enterococci clones and EFr + pPD1 led to three observations: first, 4 out of 10 clones were susceptible to bac-21 and were eliminated by EFr + pPD1 (no. 3, 4, 8 and 9); second, 6 out of 10 were immune to bac-21 killing (no. 1, 2, 5, 6, 7 and 10); third, probing for *bacA* provided evidence for pPD1-containing indigenous enterococci transconjugants in 4 out of the 6 immune indigenous enterococci clones. The two clones that failed to conjugate and were resistant to bac-21 killing in the mixed culture experiment were also resistant to EFr + pPD1 on bacteriocin assay plates. The mechanism for resistance of these two clones is not clear; however, they might have harboured cross-resistance traits. Error bars represent s.e.m.

Crystal structure of the 500-kDa yeast acetyl-CoA carboxylase holoenzyme dimer

Jia Wei¹ & Liang Tong¹

Acetyl-CoA carboxylase (ACC) has crucial roles in fatty acid metabolism and is an attractive target for drug discovery against diabetes, cancer and other diseases^{1–6}. *Saccharomyces cerevisiae* ACC (ScACC) is crucial for the production of very-long-chain fatty acids and the maintenance of the nuclear envelope^{7,8}. ACC contains biotin carboxylase (BC) and carboxyltransferase (CT) activities, and its biotin is linked covalently to the biotin carboxyl carrier protein (BCCP). Most eukaryotic ACCs are 250-kilodalton (kDa), multi-domain enzymes and function as homodimers and higher oligomers. They contain a unique, 80-kDa central region that shares no homology with other proteins. Although the structures of the BC, CT and BCCP domains and other biotin-dependent carboxylase holoenzymes are known^{1,9–14}, there is currently no structural information on the ACC holoenzyme. Here we report the crystal structure of the full-length, 500-kDa holoenzyme dimer of ScACC. The structure is remarkably different from that of the other biotin-dependent carboxylases. The central region contains five domains and is important for positioning the BC and CT domains for catalysis. The structure unexpectedly reveals a dimer of the BC domain and extensive conformational differences compared to the structure of the BC domain alone, which is a monomer. These structural changes reveal why the BC domain alone is catalytically inactive and define the molecular mechanism for the inhibition of eukaryotic ACC by the natural product soraphen A^{15,16} and by phosphorylation of a Ser residue just before the BC domain core in mammalian ACC. The BC and CT active sites are separated by 80 Å, and the entire BCCP domain must translocate during catalysis.

The primary sequences of the single-chain, multi-domain eukaryotic ACCs can be divided into three regions of roughly equal sizes. The amino-terminal region (residues 1–795 in ScACC, Fig. 1a) contains the BC and BCCP domains, with possibly a BT (BC–CT interaction) domain between them, as observed in the structures of propionyl-CoA carboxylase (PCC)¹¹ and 3-methylcrotonyl-CoA carboxylase (MCC)¹². The carboxy-terminal region (residues 1492–2233) contains the N and C domains of CT¹⁷. The structure and function of the central region (residues 796–1491) are currently not known. It is not as well conserved among the eukaryotic ACCs (Extended Data Figs 1–3). BC catalyses the MgATP-dependent carboxylation of the N1' atom of biotin (Extended Data Fig. 4). The carboxybiotin (and BCCP) then translocates to the CT active site, where the substrate acetyl-CoA is carboxylated.

We expressed in *Escherichia coli* the 250-kDa ScACC (residues 22–2233) and determined its crystal structure at 3.2 Å resolution (Extended Data Table 1, Extended Data Fig. 4). The structure of the ScACC holoenzyme is remarkably different from that of the other biotin-dependent carboxylases^{9–14}. The 500-kDa holoenzyme dimer obeys two-fold symmetry, and its overall structure is shaped like a quarter of a disk (Fig. 1b), with a radius of ~140 Å and thickness of ~120 Å (Fig. 1c), although there is a large channel measuring ~30 Å across through the centre of the holoenzyme (Extended Data Fig. 5). A BC domain dimer (Fig. 1d) is located near the centre of the

disk, while the CT domain dimer (Fig. 1e) forms a part of the edge of the disk. A BCCP domain is positioned near the centre of each face of the holoenzyme, and its biotin is located in the CT active site (Fig. 1b).

We have also determined the structure at 3.1 Å resolution of the holoenzyme where the BCCP domain is not biotinylated (Extended Data Table 1). The overall structure of this dimer is essentially the same as the other structure, with r.m.s. distance of 0.45 Å for their 3,928 equivalent C α atoms, although the BCCP domain is disordered in the absence of biotinylation. Earlier studies showed that biotinylation stabilizes *E. coli* BCCP^{18,19}. This structure of the holoenzyme will not be discussed further here.

The overall structures of the two protomers of the holoenzyme are similar, with r.m.s. distance of 1.1 Å for 1,862 equivalent C α atoms located within 3 Å of each other after superposition. On the other hand, the r.m.s. distance is only 0.76 Å if the CT domains are superposed, and differences in the orientation and position of the other domains are visible, especially for the BC domain, which is located furthest from CT (Extended Data Fig. 4). Within CT, conformational differences in the small inserted domain¹⁷ are observed (Extended Data Fig. 5), likely to be linked to differences in the position of BCCP in the two protomers, as the insert domain has direct contacts with BCCP.

The structure reveals that the central region of ScACC contains five domains, which we have named ACC central (AC) domains AC1 through AC5, giving a total of 10 major domains for each ScACC protomer (Fig. 1a). Domains AC1, AC2 and AC3 are all helical (Fig. 1f, Extended Data Fig. 2). AC1 contains three pairs of anti-parallel helices as well as inserts of a four-helical bundle (domain AC2, helices $\alpha 3$ – $\alpha 6$) and a helical hairpin ($\alpha 8$ – $\alpha 9$, Extended Data Fig. 5). AC3 is also a four-helical bundle but it has no interactions with AC1 and AC2, and instead is positioned between AC4 and AC5, mediating interactions between them. Domain AC4 is located at the end of the disk edge, and is separated from the AC4 domain of the other protomer by ~200 Å (Fig. 1b). Unexpectedly, the structure shows that AC4 and AC5 have similar backbone folds, consisting of a twisted β -sheet flanked on one face by helices (Extended Data Figs 4, 6). The r.m.s. distance is 2.9 Å for their equivalent C α atoms, but the sequence identity is only 12%. This backbone fold has weak similarity to a part of formamidase²⁰ (Extended Data Fig. 6) and several other enzymes, but the Z score is below 6.5 and the sequence identity is less than 9%²¹. The active sites of these enzymes are not conserved in AC4 or AC5. Therefore, this structural similarity is unlikely to have any functional significance for ACC.

The structure confirms the presence of a BT domain in the N-terminal region of ScACC (Fig. 1a). Its structure is similar to that in PCC¹¹, with a central, long helix (Extended Data Fig. 4) surrounded by an eight-stranded anti-parallel β -barrel (Extended Data Fig. 5). The domain helps to mediate interactions between the BC and CT domains, as there are no direct contacts between them in the holoenzyme (Fig. 1b). Within each protomer, the last three strands of the BT domain β -barrel ($\beta 27$ – $\beta 29$) faces the BC domain (Figs 1b, 2a), while the C-terminal end of the long helix and the following loop connecting

¹Department of Biological Sciences, Columbia University, New York, New York 10027, USA.

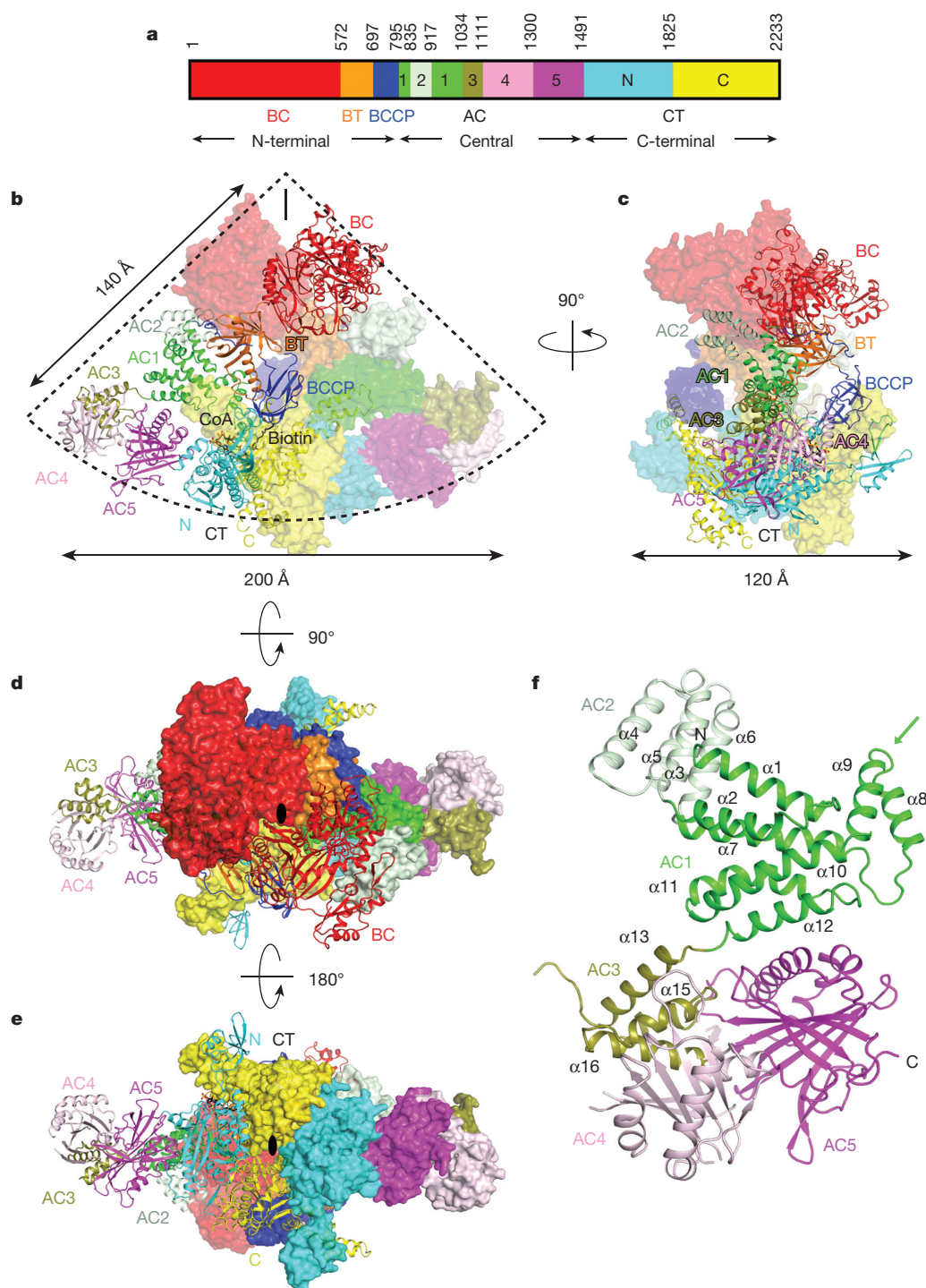


Figure 1 | Crystal structure of the 500-kDa yeast acetyl-CoA carboxylase (ScACC) holoenzyme dimer. **a**, Domain organization of ScACC. The three regions of the sequence are also indicated. AC, ACC central. **b**, Overall structure of ScACC holoenzyme dimer. One protomer is shown as ribbons while the other is shown only as a surface for clarity, both coloured according to **a**. The two-fold axis of the dimer is vertical (black line). **c–e**, Overall structure of

ScACC holoenzyme, viewed from the side (**c**), down the BC domain dimer (**d**), and down the CT domain dimer (**e**). The two-fold axis is indicated with the black oval. **f**, Structure of the five domains (AC1–AC5) in the central region of ScACC. The arrow points to the helical hairpin insert of AC1. The structure figures were produced with PyMOL (<http://www.pymol.org>).

to the first strand (the hook¹¹) is flanked by the helical hairpin insert of AC1 ($\alpha 8$ – $\alpha 9$) on one side and an inserted segment of a long loop between two β -strands ($\beta 4A$ and $\beta 4B$) in the C domain of CT on the other side (Fig. 2b, Extended Data Fig. 4). In addition, a part of the long linker between the BCCP and AC1 domains has hydrophobic interactions with the top of one side of the BT domain β -barrel (Fig. 1b, Extended Data Fig. 5).

A total of $\sim 9,000 \text{ \AA}^2$ of the surface area of each protomer is buried in the dimer interface, predominantly from the BC ($1,200 \text{ \AA}^2$, Extended Data Fig. 7) and CT ($5,800 \text{ \AA}^2$, Extended Data Fig. 8) domain dimers. The BC domain contributes an additional 900 \AA^2 through contacts with the BT domain (500 \AA^2 , Fig. 2a), AC1 (80 \AA^2) and AC2 (230 \AA^2 , Fig. 2c) of the other protomer. The BCCP domain buries $\sim 400 \text{ \AA}^2$ in the CT active site, where it contacts the C domain of the

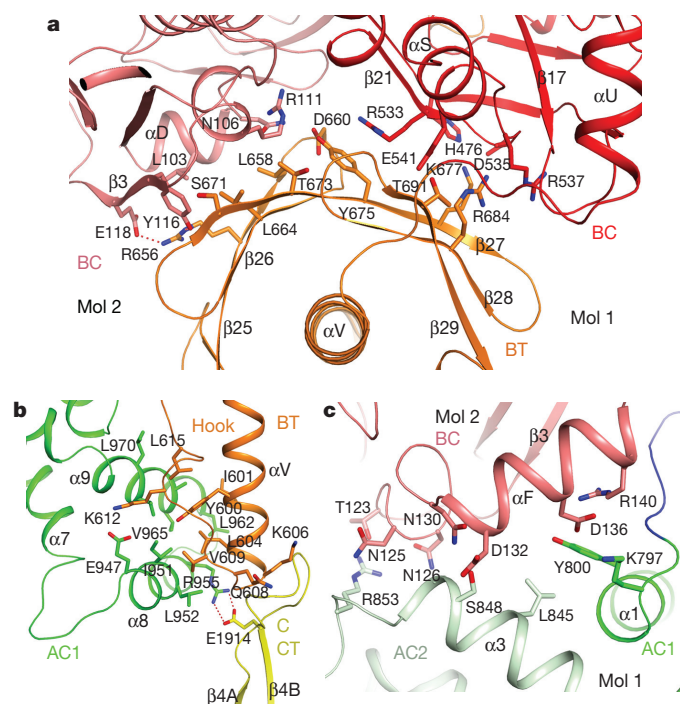


Figure 2 | Interactions among the domains in the ScACC holoenzyme.

a, The BT domain contacts the BC domain dimer. Side chains of residues in the interfaces between the BT domain (orange) and the BC domain of the same protomer (red, Mol 1) and the BC domain of the other protomer (salmon, Mol 2) are shown as stick models. **b**, Interactions between the hook of the BT domain (orange) and the helical hairpin insert of AC1 domain ($\alpha 8$ and $\alpha 9$, green) and the $\beta 4A$ – $\beta 4B$ loop from the C domain of CT (yellow). **c**, Interactions between domains AC1 (green) and AC2 (light green) of one protomer with the BC domain (salmon) of the other protomer.

other protomer, but it is expected to translocate to the BC active site during catalysis (see later).

A major surprise from the structure is the observation of a BC domain dimer (Fig. 1d), because the BC domain alone is consistently a monomer and catalytically inactive based on earlier studies^{15,16,22,23}. Moreover, the organization of this BC domain dimer is similar to that of the BC subunit dimer of *E. coli* ACC^{24–26} (Fig. 3a), with a mostly hydrophilic interface (Extended Data Fig. 7). However, the structure of BC domain alone is incompatible with such a dimer due to steric clashes between the two molecules¹⁶ (Extended Data Fig. 7). Large conformational changes for residues in the dimer interface are therefore necessary for the formation of this dimer, primarily involving the β -strands and connecting loops in the C sub-domain of BC (Fig. 3b). Especially, strand $\beta 18$ moves by ~ 8 Å, taking it out of the central β -sheet of the C-domain (Fig. 3c, Supplementary Video 1). $\beta 18$ instead forms a β -sheet with a new strand, $\beta 21$, which is not present in the structure of BC domain alone. Residues in the $\beta 17$ – $\beta 18$ loop move by up to 20 Å, and those in the $\beta 18$ – $\beta 19$ loop by up to 7 Å (Fig. 3c). In addition, the main chain of neighbouring strands $\beta 17$, $\beta 19$ and $\beta 20$ shifts by ~ 3 Å. The new $\beta 18$ -loop- $\beta 19$ structure is in the centre of the BC domain dimer (Fig. 3a), where the tip of this loop contacts the side chain of Trp487 in the other protomer (Extended Data Fig. 7).

The two distinct conformations of this domain explain why it is catalytically inactive on its own¹⁵ and also define the molecular mechanism for the inhibition of eukaryotic ACC by the natural product soraphen A¹⁶ and by phosphorylation of a Ser residue just before the BC domain core in mammalian ACC (Ser80 in human ACC1 and Ser222 in ACC2)²³ (Extended Data Fig. 1). As a part of the conformational change, residues in the $\beta 19$ – $\beta 20$ loop move by up to 10 Å (Fig. 3d, Supplementary Video 2). This loop in the structure of BC domain alone is likely to interfere with the binding of BCCP–biotin (Fig. 3d, Extended Data Fig. 7), based on the

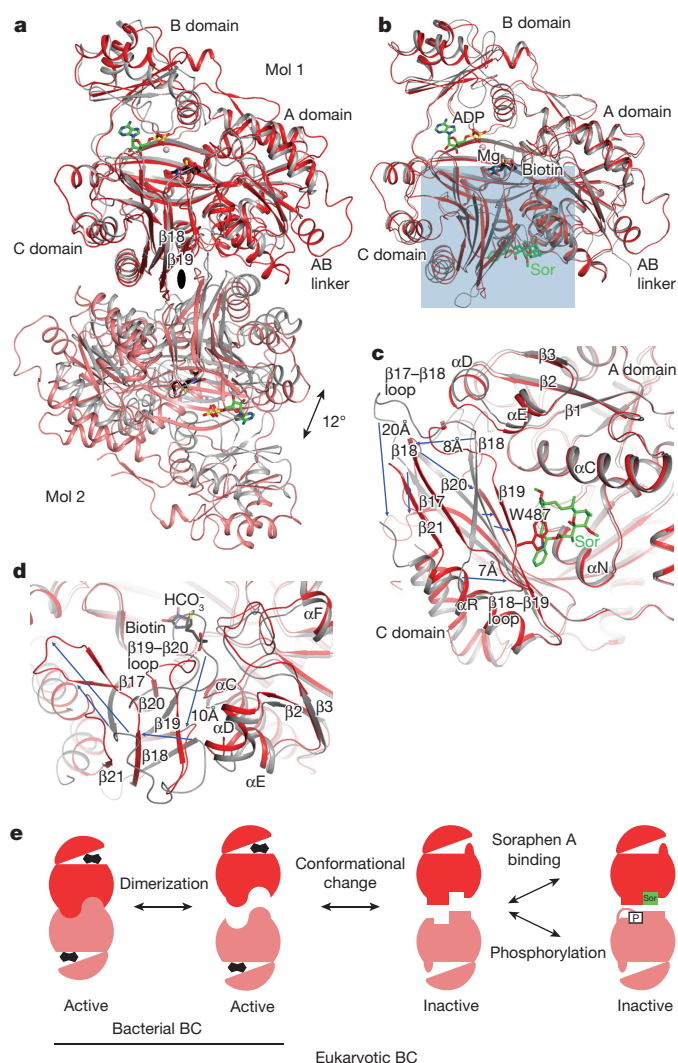


Figure 3 | A dimeric BC domain in the ScACC holoenzyme. **a**, Overlay of the BC domain dimer of ScACC (in red and salmon) with the BC subunit dimer of *E. coli* ACC (grey) in complex with ADP (green), bicarbonate (black), biotin (black), and Mg^{2+} (pink sphere)²⁶. The two molecules at the top are superposed, and the two molecules at the bottom have $\sim 12^\circ$ difference in orientation. **b**, Overlay of the BC domain of ScACC holoenzyme (in red) with the BC domain alone (in grey) in complex with soraphen A (green, labelled Sor)¹⁶. The region of large conformational differences is highlighted in light blue. The view is the same as in **a**. **c**, Detailed view of the conformational changes in the dimer interface. Blue arrows indicate some of the changes from the BC domain alone (grey) to the BC domain in the holoenzyme (red). **d**, Conformational changes near the biotin binding site, especially the $\beta 19$ – $\beta 20$ loop. This is coupled to changes in the $\beta 2$ to αF segment. A possible hydrogen bond between the amide linkage of biotin and the $\beta 19$ – $\beta 20$ loop is indicated with the dashed lines (red). **e**, A model for how conformational transitions in the dimer interface affect catalysis and dimerization of the eukaryotic BC domain, updated from an earlier model²⁵. Biotin is shown as the fused black pentagons, while soraphen A and phosphorylated serine are indicated by Sor and P, respectively. Bacterial BC subunit does not undergo the conformational transition, and its monomers can be catalytically active²⁵.

binding mode of biotin to *E. coli* BC²⁶. Consequently, BC domain alone is catalytically inactive because it assumes a conformation that cannot bind the BCCP–biotin substrate.

Soraphen A recognizes the conformation of isolated BC domain¹⁶, and this binding site does not exist in the BC domain dimer in the holoenzyme owing to the structural changes (Fig. 3c). For example, strands $\beta 19$ and $\beta 20$ move into the binding site, and especially the side

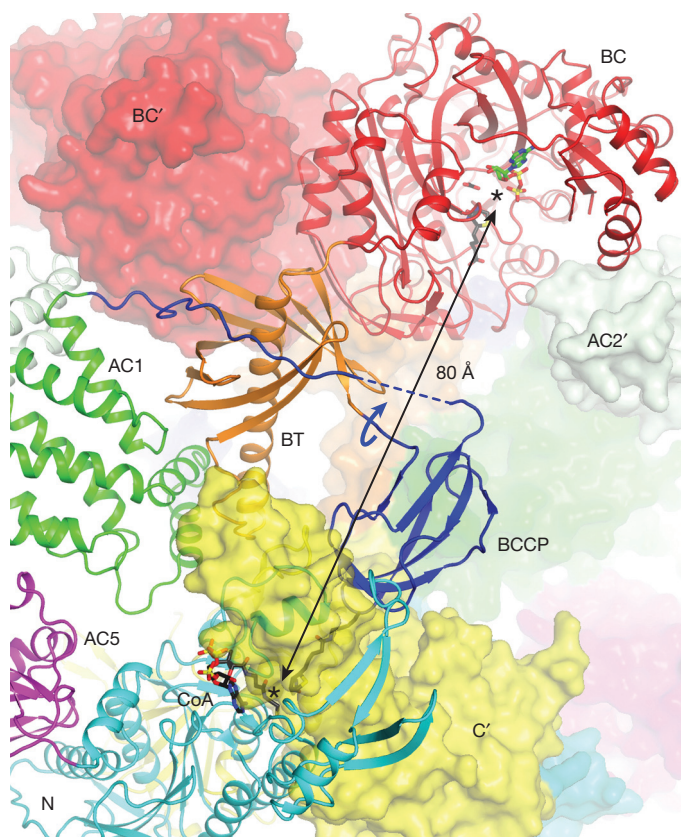


Figure 4 | Translocation of BCCP and biotin during ACC catalysis. The BC and CT active sites (black asterisks) of ScACC are separated by ~ 80 Å (arrow). A rotation of $\sim 180^\circ$ around the BT–BCCP linker (curved arrow in blue) could place the biotin into the BC active site. The prime in the labels indicates the second protomer. The binding modes of ADP (green) and biotin (black) to *E. coli* BC subunit are shown as stick models²⁶.

chain of Trp487 (B19) is in direct clash with soraphen A (Fig. 3c, Extended Data Fig. 7). Therefore, soraphen A inhibits the enzyme allosterically by stabilizing a catalytically inactive conformation of the BC domain. In the context of the holoenzyme, soraphen A binding will disrupt the formation of BC domain dimer (Fig. 1d), as the other protomer clashes with the compound as well (Extended Data Fig. 7). This could also be detrimental for catalysis as the BC domains may not be positioned correctly to accept the BCCP–biotin for carboxylation.

Upon phosphorylation, the peptide segment containing pSer222 of human ACC2 is located in the same binding site as soraphen A²³, and the Trp487 side chain in the holoenzyme structure also clashes with this segment (Extended Data Fig. 7). Therefore, phosphorylation of this Ser residue inhibits the enzyme through the same mechanism as that of soraphen A. These structural observations greatly extend a model for the inhibitory mechanism proposed earlier²⁵ (Fig. 3e). ScACC does not have an equivalent phosphorylation site in this region of the sequence.

In comparison, the structure of the CT domain dimer in the holoenzyme is essentially the same as that of the domain alone¹⁷, with r.m.s. distance of 0.61 Å among their equivalent C α atoms (Extended Data Fig. 8). The binding of both BCCP–biotin and CoA (Extended Data Fig. 4) in one of the CT active sites provides direct insights into the catalysis by this enzyme. The thiol group of CoA is 4.3 Å away from the N1' atom of biotin (Extended Data Fig. 8). Therefore, the two substrates are likely to be in the correct positions for catalysis. The position of biotin clashes with that of the compound CP-640186 (Extended Data Fig. 8), a nanomolar inhibitor of mammalian ACCs²⁷, confirming that it functions by blocking biotin binding to the CT active site²⁸.

Prior to the structure determination of the holoenzyme, we obtained the crystal structures for residues 797–1033 (domains AC1–AC2), 1036–1503 (AC3–AC5), and 569–1494 (BT, BCCP and the entire central region) (Extended Data Table 1). Comparisons of the structures of these domains alone with that of the holoenzyme reveal substantial variability in the relative positioning of the domains. Domains AC3–AC4 in the structure of AC3–AC5 alone can be readily superposed with those in the holoenzyme, but then the orientation of domain AC5 differ by 40° (Extended Data Fig. 9). Even more variability is observed in the structure of BT–BCCP–AC1–AC5 alone, illustrated by the differences in the positioning of AC1–AC2 relative to the BT domain and AC3–AC5 (Extended Data Fig. 9). Interestingly, between the two unique copies of the BT–BCCP–AC1–AC5 molecule in the crystal, one has a conformation of AC3–AC5 that is very similar to that in the holoenzyme (Extended Data Fig. 9), while the other is similar to that in AC3–AC5 alone, suggesting that domains AC3–AC5 may assume two (or more) distinct conformations.

Another discovery from the structure is that the central region has minimal contributions to the formation of the dimer. Besides the ~ 300 Å² surface area burial for AC1 and AC2 (Fig. 2c), the central region has no contacts with the other protomer (Fig. 1b). On the other hand, the structure suggests that the central region is important for maintaining the BC and CT dimers in the correct relative positions for catalysis. The CT domain dimer is sandwiched by AC5 on both sides (Fig. 1e). The BT and AC2 domains of the two protomers form a platform, keeping the BC domain dimer in place and possibly also helping with BC domain dimerization (Fig. 1b). The conformational variability observed for the domains in the central region (Extended Data Fig. 9) might have a role in regulating the activity of the holoenzyme.

The structural analysis indicates that BCCP–biotin becomes carboxylated in the BC active site of its own protomer, and then translocates to the CT active site at the dimer interface, where it contacts the C domain of the other protomer (Fig. 4). The distance between the BC and CT active sites is ~ 80 Å, indicating that the entire BCCP domain must translocate during catalysis (swinging-domain model, Extended Data Fig. 4), as has been observed in the other holoenzymes^{1,9–14}. In fact, the linker from BT to BCCP (residues 697–700) is about 45 Å from the N1' atom of biotin, and therefore a rotation of this linker by $\sim 180^\circ$ could bring the biotin into the BC active site from the CT active site (Fig. 4). The linker from BCCP to AC1 (residues 770–795) is much longer and can accommodate such a rotation, and part of this linker is disordered in the current structure.

We introduced mutations in the interfaces in the holoenzyme to assess the structural information. The mutants that could be purified migrated at the same position on a gel filtration column and had nearly the same thermal melting curves as the wild-type enzyme (data not shown). Deletion of the α -helical hairpin ($\alpha 8$ – $\alpha 9$, residues 940–972) in domain AC1 (Extended Data Fig. 2) or the $\beta 4A$ – $\beta 4B$ loop in the C domain of CT (residues 1902–1916, Extended Data Fig. 3) abolished the catalytic activity (Table 1), demonstrating their importance in anchoring the hook of the BT domain (Fig. 2b). On the other hand, mutating Gln608 in the hook, which interacts with the main chain of

Table 1 | Effects of mutations in the ScACC holoenzyme interfaces on the catalysis

Enzyme	K_m (mM)*	k_{cat} (s ^{−1})
Wild-type ScACC	0.053 ± 0.011	8.5 ± 0.4
$\Delta 940$ –972 ($\alpha 8$ – $\alpha 9$ hairpin of AC1)	No activity detected	
$\Delta 1902$ –1916 ($\beta 4A$ – $\beta 4B$ loop of CT)	No activity detected	
$\Delta 836$ –918 (AC2)	No expression	
K73E	Very low activity	
R76E	No activity detected	
Y83A	0.074 ± 0.020	1.6 ± 0.1
W487A	Very low activity	
Q608R	0.11 ± 0.02	16 ± 1.0
R656E	0.15 ± 0.07	5.0 ± 0.6

* The errors were obtained from fitting data to the Michaelis–Menten equation.

the β 4A– β 4B loop (Fig. 2b), had little effect on catalysis. Similarly, the R656E mutation at the edge of the BT–BC interface (Fig. 2a) had little effect, suggesting that these single-site mutations are not sufficient to disrupt the holoenzyme or that these regions of contact do not contribute significantly to the interactions.

We introduced the K73E, R76E and W487A mutations in the BC domain dimer interface (Extended Data Fig. 7), and all three essentially abolished the catalytic activity (Table 1). The mutations probably disrupted the BC domain dimer, and the domain changed to the other conformation that is incompatible with catalysis. The K73E and R76E mutants are equivalent to the R16E and R19E mutants of *E. coli* BC subunit that we characterized earlier^{25,29}. Those mutations greatly destabilized the dimer, but had only a small effect (about three-fold) on the catalytic activity of the BC subunit alone, probably because the *E. coli* BC monomer does not undergo the conformational transition as observed here for the eukaryotic BC (Fig. 3e). However, the R19E mutation had a much larger effect *in vivo*, probably in the context of the *E. coli* ACC holoenzyme³⁰.

Overall, our studies have produced the first structural information on the 500-kDa yeast ACC holoenzyme dimer. This structure is likely to have relevance for other eukaryotic ACCs, especially the human ACC holoenzymes, as they share 45% sequence identity with yeast ACC. The structures of the BC and CT domains of human and yeast ACCs are highly similar and consistent with their strong sequence conservation (Extended Data Figs 1, 3). Although the sequence conservation for the central region is weaker, the secondary structure elements in yeast ACC are predicted to be present in human ACC as well, and residues in these secondary structure elements are more conserved than those in the loops (Extended Data Fig. 2). Therefore, the overall structure of human ACC holoenzyme is likely to be similar to that of yeast ACC.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 22 March; accepted 4 August 2015.

Published online 12 October 2015.

- Tong, L. Structure and function of biotin-dependent carboxylases. *Cell. Mol. Life Sci.* **70**, 863–891 (2013).
- Waldrop, G. L., Holden, H. M. & St Maurice, M. The enzymes of biotin dependent CO₂ metabolism: what structures reveal about their reaction mechanisms. *Protein Sci.* **21**, 1597–1619 (2012).
- Cronan, J. E. Jr & Waldrop, G. L. Multi-subunit acetyl-CoA carboxylases. *Prog. Lipid Res.* **41**, 407–435 (2002).
- Polyak, S. W., Abell, A. D., Wilce, M. C. J., Zhang, L. & Booker, G. W. Structure, function and selective inhibition of bacterial acetyl-CoA carboxylase. *Appl. Microbiol. Biotechnol.* **93**, 983–992 (2012).
- Abramson, H. N. The lipogenesis pathway as a cancer target. *J. Med. Chem.* **54**, 5615–5638 (2011).
- Wakil, S. J. & Abu-Elheiga, L. A. Fatty acid metabolism: target for metabolic syndrome. *J. Lipid Res.* **50** (Suppl), S138–S143 (2009).
- Schneider, R. et al. A yeast acetyl coenzyme A carboxylase mutant links very-long-chain fatty acid synthesis to the structure and function of the nuclear membrane-pore complex. *Mol. Cell. Biol.* **16**, 7161–7172 (1996).
- Hoja, U., Wellein, C., Greiner, E. & Schweizer, E. Pleiotropic phenotype of acetyl-CoA-carboxylase-defective yeast cells. *Eur. J. Biochem.* **254**, 520–526 (1998).
- St Maurice, M. et al. Domain architecture of pyruvate carboxylase, a biotin-dependent multifunctional enzyme. *Science* **317**, 1076–1079 (2007).
- Xiang, S. & Tong, L. Crystal structures of human and *Staphylococcus aureus* pyruvate carboxylase and molecular insights into the carboxyltransfer reaction. *Nature Struct. Mol. Biol.* **15**, 295–302 (2008).
- Huang, C. S. et al. Crystal structure of the $\alpha_6\beta_6$ holoenzyme of propionyl-coenzyme A carboxylase. *Nature* **466**, 1001–1005 (2010).
- Huang, C. S., Ge, P., Zhou, Z. H. & Tong, L. An unanticipated architecture of the 750-kDa $\alpha_6\beta_6$ holoenzyme of 3-methylcrotonyl-CoA carboxylase. *Nature* **481**, 219–223 (2012).
- Fan, C., Chou, C.-Y., Tong, L. & Xiang, S. Crystal structure of urea carboxylase provides insights into the carboxyltransfer reaction. *J. Biol. Chem.* **287**, 9389–9398 (2012).
- Tran, T. H. et al. Structure and function of a single-chain, multi-domain long-chain acyl-CoA carboxylase. *Nature* **518**, 120–124 (2015).
- Weatherly, S. C., Volrath, S. L. & Elich, T. D. Expression and characterization of recombinant fungal acetyl-CoA carboxylase and isolation of a soraphen-binding domain. *Biochem. J.* **380**, 105–110 (2004).
- Shen, Y., Volrath, S. L., Weatherly, S. C., Elich, T. D. & Tong, L. A mechanism for the potent inhibition of eukaryotic acetyl-coenzyme A carboxylase by soraphen A, a macrocyclic polyketide natural product. *Mol. Cell* **16**, 881–891 (2004).
- Zhang, H., Yang, Z., Shen, Y. & Tong, L. Crystal structure of the carboxyltransferase domain of acetyl-coenzyme A carboxylase. *Science* **299**, 2064–2067 (2003).
- Chapman-Smith, A., Forbes, B. E., Wallace, J. C. & Cronan, J. E. Jr. Covalent modification of an exposed surface turn alters the global conformation of the biotin carrier domain of *Escherichia coli* acetyl-CoA carboxylase. *J. Biol. Chem.* **272**, 26017–26022 (1997).
- Solbiati, J., Chapman-Smith, A. & Cronan, J. E. Jr. Stabilization of the biotinoyl domain of *Escherichia coli* acetyl-CoA carboxylase by interactions between the attached biotin and the protruding “thumb” structure. *J. Biol. Chem.* **277**, 21604–21609 (2002).
- Hung, C. L. et al. Crystal structure of *Helicobacter pylori* formamidase AmfF reveals a cysteine-glutamate-lysine catalytic triad. *J. Biol. Chem.* **282**, 12220–12229 (2007).
- Holm, L., Kääriäinen, S., Rosenström, P. & Schenkel, A. Searching protein structure databases with DALI Lite v.3. *Bioinformatics* **24**, 2780–2781 (2008).
- Raymer, B. et al. Synthesis and characterization of a BODIPY-labeled derivative of Soraphen A that binds to acetyl-CoA carboxylase. *Bioorg. Med. Chem. Lett.* **19**, 2804–2807 (2009).
- Cho, Y. S. et al. Molecular mechanism for the regulation of human ACC2 through phosphorylation by AMPK. *Biochem. Biophys. Res. Commun.* **391**, 187–192 (2010).
- Waldrop, G. L., Rayment, I. & Holden, H. M. Three-dimensional structure of the biotin carboxylase subunit of acetyl-CoA carboxylase. *Biochemistry* **33**, 10249–10256 (1994).
- Shen, Y., Chou, C.-Y., Chang, G.-G. & Tong, L. Is dimerization required for the catalytic activity of bacterial biotin carboxylase? *Mol. Cell* **22**, 807–818 (2006).
- Chou, C.-Y., Yu, L. P. C. & Tong, L. Crystal structure of biotin carboxylase in complex with substrates and implications for its catalytic mechanism. *J. Biol. Chem.* **284**, 11690–11697 (2009).
- Harwood, H. J. Jr et al. Isozyme-nonselective N-substituted bipiperidylcarboxamide acetyl-CoA carboxylase inhibitors reduce tissue malonyl-CoA concentrations, inhibit fatty acid synthesis, and increase fatty acid oxidation in cultured cells and in experimental animals. *J. Biol. Chem.* **278**, 37099–37111 (2003).
- Zhang, H., Tweel, B., Li, J. & Tong, L. Crystal structure of the carboxyltransferase domain of acetyl-coenzyme A carboxylase in complex with CP-640186. *Structure* **12**, 1683–1691 (2004).
- Chou, C.-Y. & Tong, L. Structural and biochemical studies on the regulation of biotin carboxylase by substrate inhibition and dimerization. *J. Biol. Chem.* **286**, 24417–24425 (2011).
- Smith, A. C. & Cronan, J. E. Dimerization of the bacterial biotin carboxylase subunit is required for acetyl coenzyme A carboxylase activity *in vivo*. *J. Bacteriol.* **194**, 72–78 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank M. Bush, C.-Y. Chou, Y. Shen, L. Yu and H. Zhang for carrying out initial studies in this project; R. Jackimowicz, B. Nolan, N. Whalen, A. Heroux and H. Robinson for access to the X29A and X25 beamlines at the NSLS; S. Banerjee, K. Perry, R. Rajashankar, J. Schuermann, N. Sukumar for access to NE-CAT 24-C and 24-E beamlines at the Advanced Photon Source; W. Rice and E. Eng at New York Structural Biology Center for assistance with electron microscopy. The in-house X-ray diffraction instrument was purchased with an NIH grant to L.T. (S100D012018). This research was supported in part by a grant from the NIH (R01DK067238) to L.T.

Author Contributions J.W. carried out protein expression, purification, crystallization, data collection, structure determination and refinement, site-directed mutagenesis and enzymatic assays. L.T. initiated the project, supervised the entire research, and analysed the results. J.W. and L.T. wrote the paper.

Author Information Coordinates and structure factors have been deposited in the Protein Data Bank under accession numbers 5CS0 for AC1–AC2, 5CS4 for AC3–AC5, 5CSA for BT–BCCP–AC1–AC5, 5CSK for unbiotinylated ACC, and 5CSL for ACC holoenzyme. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to L.T. (ltong@columbia.edu).

METHODS

Protein expression and purification. The N-terminal segment of the central region of *Saccharomyces cerevisiae* ACC (ScACC, gene *ACC1*, residues 797–1033, AC1–AC2 domains) was expressed at 20 °C in *E. coli* BL21(DE3) Star cells for native protein or B384(DE3) cells for selenomethionyl protein, in the presence of a chaperone plasmid pG-KJE8 (TaKaRa). The recombinant protein carried a C-terminal His-tag and was purified by Ni-NTA (Qiagen) and gel filtration chromatography (Sephacryl S-300, GE Healthcare) in buffer A (20 mM Tris (pH 7.5), 300 mM NaCl, and 2 mM DTT).

The C-terminal segment of the central region of ScACC (residues 1036–1503, AC3–AC5 domains) was expressed in *E. coli* B834(DE3) cells at 25 °C for selenomethionyl protein and purified following the same protocol. The gel filtration buffer contained 10 mM rather than 2 mM DTT.

The segment containing the BT–BCCP–AC1–AC5 domains of ScACC (residues 569–1494) was expressed in BL21(DE3) Rosetta cells at 24 °C. The recombinant protein, with a C-terminal His-tag, was purified by Ni-NTA and gel filtration chromatography in buffer A.

Full-length ScACC (residues 1–2233) was constructed into pET28a (Novagen) by sewing together two PCR fragments with a 900-nt overlap. For structure determination, the segment containing residues 22–2233, with a C-terminal His-tag, was expressed in *E. coli* BL21(DE3) Star cells at 25 °C. Protein expression for all the different ScACC segments was driven by the *trp* promoter and was induced with 3-indoleacrylic acid. However, the endogenous *E. coli* biotin-protein ligase (BPL, also known as BirA) was not able to biotinylate the ScACC. The ScBPL gene was amplified from the genome and inserted into the pCDFDuet-1 vector (Novagen) multiple cloning site 2 without any affinity tag. Co-expression of ScBPL and the inclusion of 20 mg l^{−1} biotin in the media allowed complete biotinylation of ScACC, confirmed by an avidin shift assay. The protein was purified by Ni-NTA and gel filtration chromatography in buffer A. Typically 0.5 mg of biotinylated ScACC (or 3 mg of unbiotinylated ScACC) could be purified from 12 l of culture.

Protein crystallization. Crystals were obtained at 20 °C using the sitting-drop vapour diffusion method. Native and selenomethionyl crystals of the AC1–AC2 domains were obtained using a precipitant solution of 90 mM Bis-tris propane and 60 mM citric acid (pH 6.4), and 20% (w/v) PEG3350. The protein concentration was 12 mg ml^{−1}, and the crystals took 3 weeks to reach full size.

Selenomethionine-substituted crystals of the AC3–AC5 domains were obtained after 2 days. The protein concentration was 3.6 mg ml^{−1}, and the precipitant solution contained 80 mM HEPES (pH 7.5), 4% (v/v) MPD, 8 mM sodium citrate, 4% (v/v) glycerol, and 40 mM NDSB-201.

Native crystals of the BT–BCCP–AC1–AC5 domains were obtained after 4 days. The protein concentration was 15 mg ml^{−1}, and the precipitant solution contained 80 mM HEPES (pH 7.5), 9.6% (w/v) PEG6000, 1.6% (v/v) MPD, 60 mM sodium citrate, and 80 mM NaI.

Full-length ACC protein was incubated with 3.3 mM acetyl-CoA and 3.3 mM Mg-ADP for 30 min on ice before crystallization. Crystals were obtained after 2 weeks. The protein concentration was 5 mg ml^{−1}, and the precipitant solution contained 14% (w/v) PEG3350, 4% (v/v) tert-butanol, and 0.2 M sodium citrate.

Glycerol was used as the cryo-protectant and all crystals were flash frozen in liquid nitrogen for data collection at 100 K.

Data collection and structure determination. X-ray diffraction data of AC1–AC2 domains were collected on an ADSC Q315 CCD at the X29A beamline of the National Synchrotron Light Source (NSLS). The diffraction images were processed with the HKL program³¹. The crystal belonged to space group *P*₆₅ with cell parameters of *a* = *b* = 117.6 Å, and *c* = 73.8 Å. There is a domain-swapped dimer of the protein in the asymmetric unit. A selenomethionyl single-wavelength anomalous diffraction (SAD) data set was collected to 3.0 Å resolution (wavelength 0.979 Å) and a native data set to 2.5 Å resolution (wavelength 1.075 Å). Five Se atoms were located with program Solve³² and used for phasing with program Phenix³³. The phase information was then extended to 2.5 Å with solvent flattening, histogram matching and two-fold non-crystallographic symmetry (NCS) averaging using the program DM in CCP4³⁴. The atomic model was built into the electron density map manually with the program Coot³⁵. Structure refinement was performed with CNS³⁶ and Refmac³⁷.

A selenomethionyl SAD data set to 3.2 Å resolution of AC3–AC5 domains was collected at the X29A beamline (wavelength 0.979 Å). The crystal belonged to space group *P*₂₁, with cell parameters of *a* = 56.8 Å, *b* = 93.3 Å, *c* = 111.1 Å, and β = 100.6°. There are two molecules in the asymmetric unit. Five Se atoms in each molecule were located and employed for phasing with program Phenix. An atomic model was manually built into the electron density map with program Coot, and the structure was refined with Refmac5.

A native diffraction data set to 3.0 Å resolution of BT–BCCP–AC1–AC5 domains was collected at the X29A beamline (wavelength 1.075 Å). The crystal belonged to space group *P*₂₁, with cell parameters of *a* = 93.3 Å, *b* = 149.7 Å, *c* = 95.4 Å, and β = 118.4°. There is a dimer of the protein in the asymmetric unit. The structures of AC1–AC2 and AC3–AC5 domains were used as the search models to solve the structure by molecular replacement with the program Phaser³⁸. The BT and BCCP domains were manually built into the electron density map.

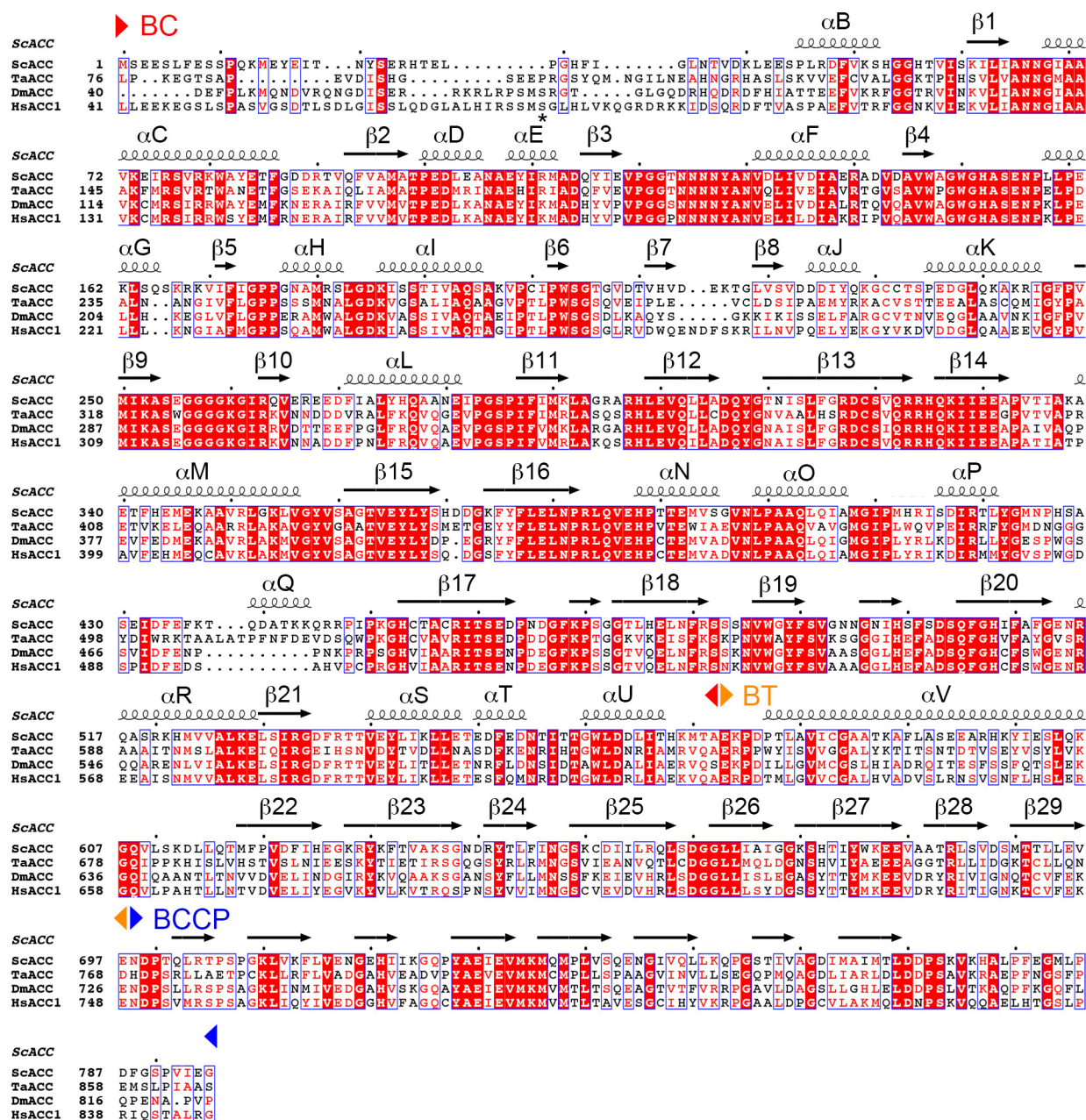
A native diffraction data set to 3.1 Å resolution of residues 22–2233 of unbiotinylated ScACC was collected on a Pilatus 6M detector at the X25 beamline of NSLS (wavelength 1.100 Å). There is a dimer of ScACC in the asymmetric unit. The structure was solved by molecular replacement with the program Phaser. The structures of AC1–AC2, AC3–AC5, and BT domains reported here and previously published structures of yeast BC¹⁶ and CT domains¹⁷ were used as the search models. However, no electron density for the BCCP domain was observed based on the crystallographic analysis.

A native diffraction data set to 3.2 Å resolution of biotinylated ScACC was collected at the X25 beamline. The crystal belonged to space group *P*₄₃₂, with cell parameters of *a* = *b* = 159.9 Å, and *c* = 614.4 Å, isomorphous to that of the unbiotinylated ScACC. The final atomic model was built with Coot and refined with Refmac5. NCS restraints were used during the refinement. Crystals of full-length ScACC (residues 1–2233) did not diffract beyond 8 Å resolution even after extensive efforts. The geometry of the final model was validated with MolProbity³⁹.

Mutagenesis and kinetic assays. Site-specific and deletion mutations were introduced with the QuikChange kit (Agilent) and sequenced for confirmation. In deletion mutants, residues 940–972 of the α-helical hairpin (α8–α9) in domain AC1 and residues 836–918 in domain AC2 were replaced by a (GS)₃ linker, respectively, while residues 1902–1916 of the β4A–β4B loop in the C domain of CT were replaced by a (GS)₂ linker.

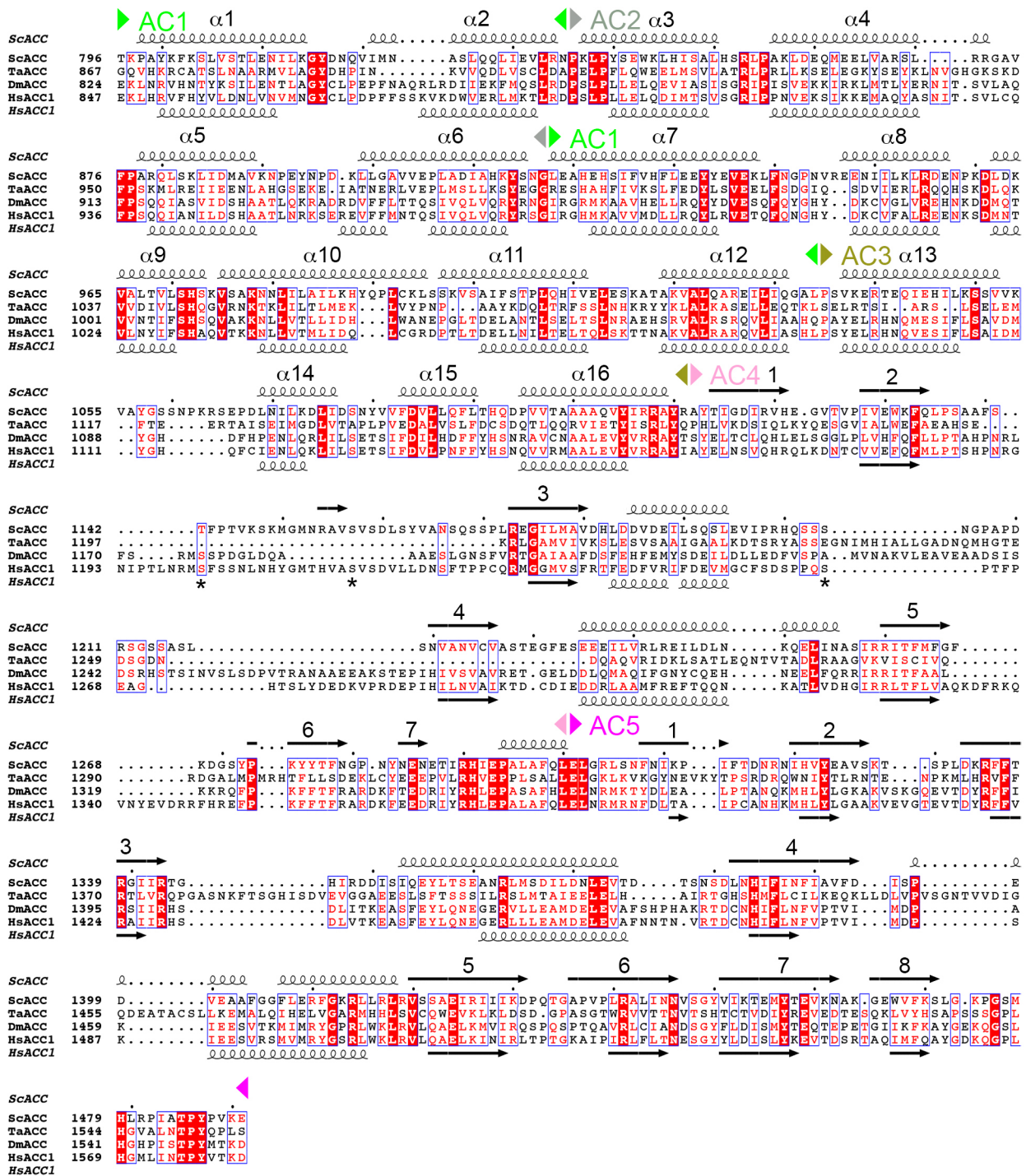
The catalytic activity of ACC was determined using a coupled enzyme assay, converting the hydrolysis of ATP to the disappearance of NADH⁴⁰. The reaction mixture contained 100 mM HEPES (pH 7.5), 8 mM MgCl₂, 40 mM KHCO₃, 200 mM KCl, 0.2 mM NADH, 0.5 mM phosphoenolpyruvate, 0.5 mM ATP, 6 units of lactate dehydrogenase (Sigma), 4 units of pyruvate kinase, 100 nM ACC and various concentrations of acetyl-CoA. The absorbance at 340 nm was monitored for 60 s.

- Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
- Terwilliger, T. C. SOLVE and RESOLVE: automated structure solution and density modification. *Methods Enzymol.* **374**, 22–37 (2003).
- Adams, P. D. et al. PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr. D* **58**, 1948–1954 (2002).
- Collaborative Computational Project, Number 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).
- Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
- Brünger, A. T. et al. Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
- Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D* **53**, 240–255 (1997).
- McCoy, A. J. et al. Phaser crystallographic software. *J. Appl. Cryst.* **40**, 658–674 (2007).
- Chen, V. B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66**, 12–21 (2010).
- Blanchard, C. Z., Lee, Y. M., Frantom, P. A. & Waldrop, G. L. Mutations at four active site residues of biotin carboxylase abolish substrate-induced synergism by biotin. *Biochemistry* **38**, 3393–3400 (1999).
- Gouet, P., Courcelle, E., Stuart, D. I. & Métz, F. ESPript: analysis of multiple sequence alignments in PostScript. *Bioinformatics* **15**, 305–308 (1999).



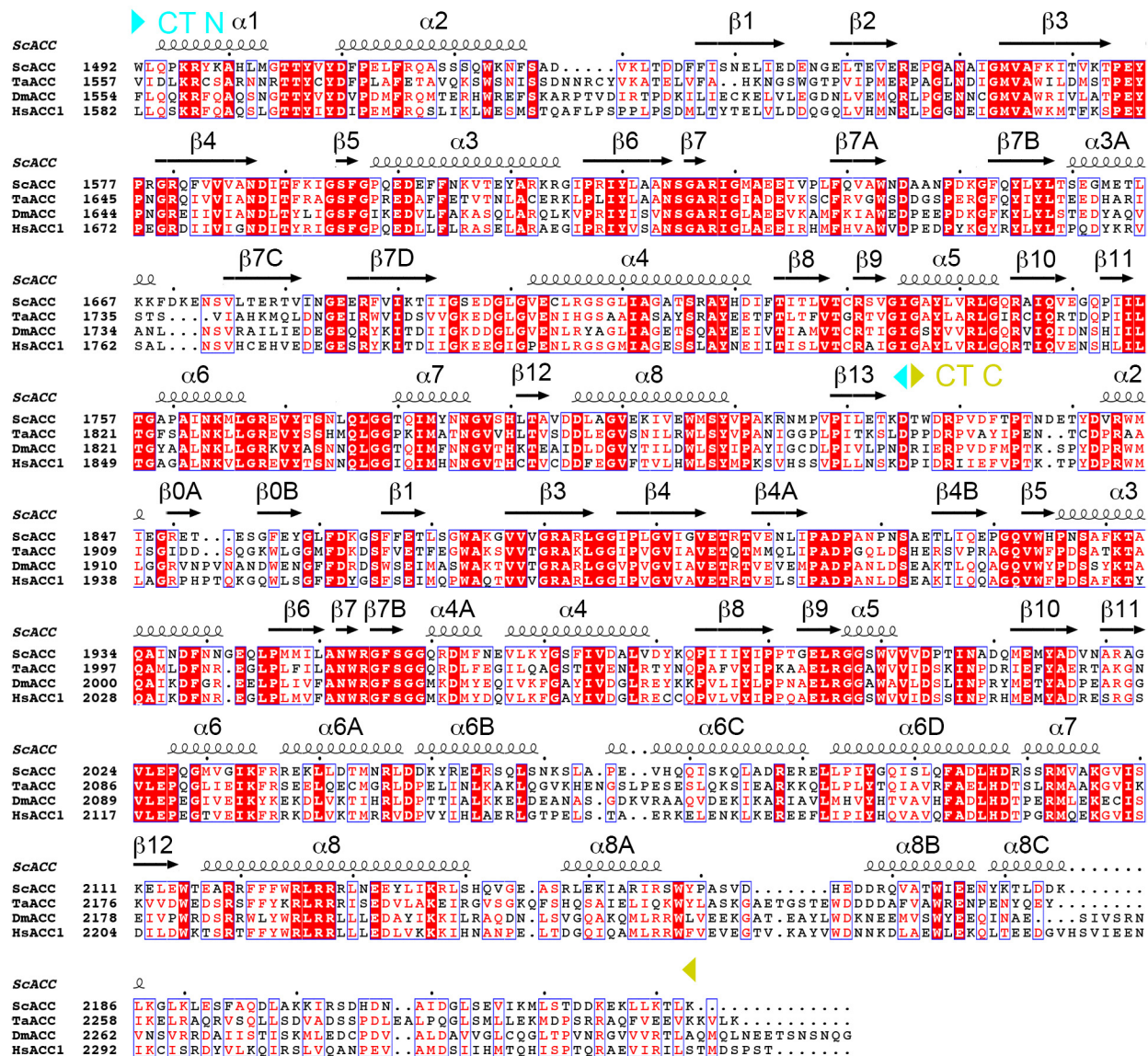
Extended Data Figure 1 | Sequence alignment of the N-terminal region of eukaryotic, single-chain ACCs. The region includes BC, BT and BCCP domains (indicated). The secondary structure elements in the ScACC holoenzyme structure are shown. The site of phosphorylation in HsACC1

(Ser80) is indicated with a star. This site does not exist in ScACC. ScACC, *Saccharomyces cerevisiae* ACC; TaACC, *Triticum aestivum* (wheat) ACC; DmACC, *Drosophila melanogaster* ACC; HsACC1, *Homo sapiens* ACC1. Modified from an output from ESPrnt⁴¹.

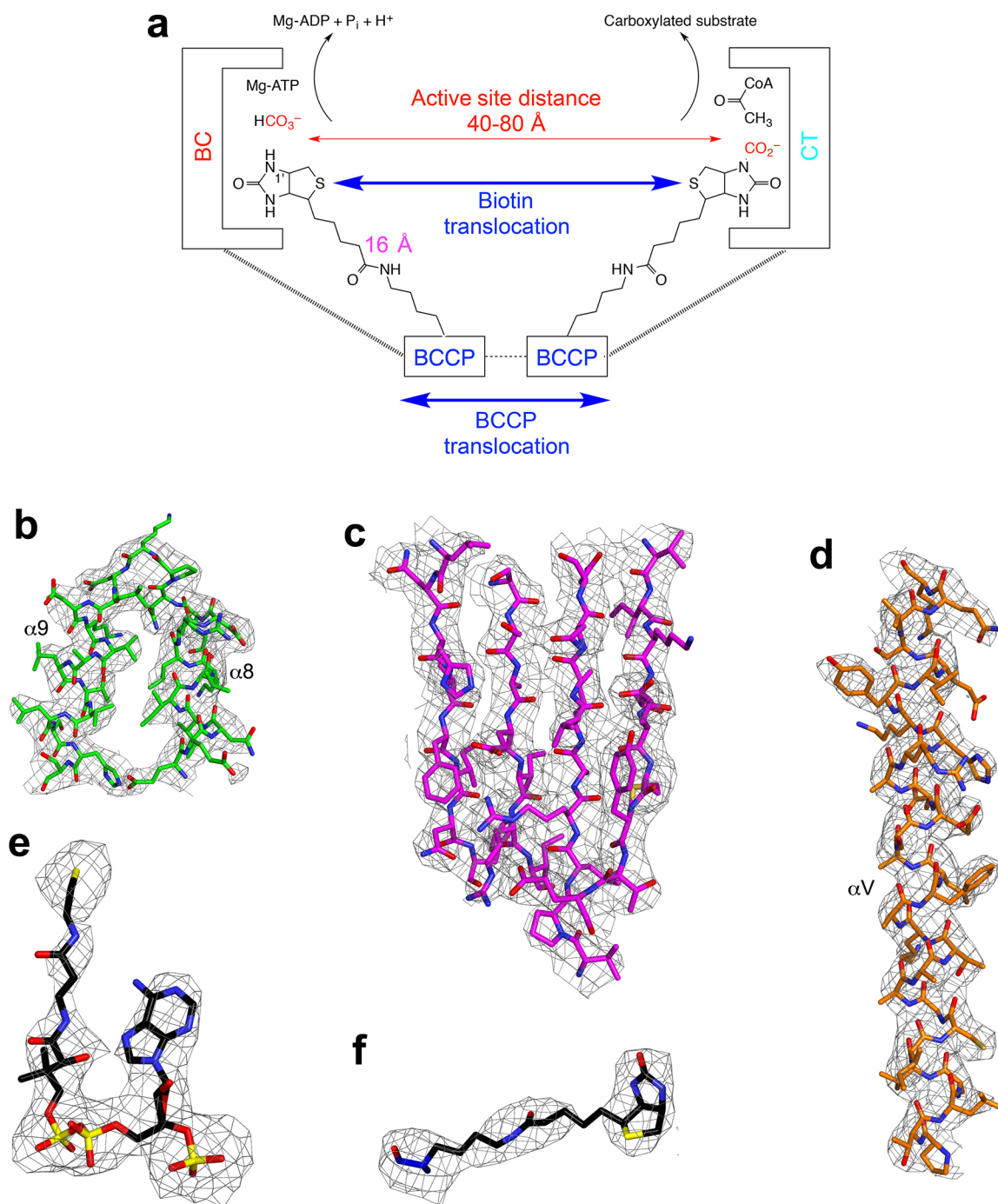


Extended Data Figure 2 | Sequence alignment of the central region of eukaryotic, single-chain ACCs. The AC1–AC5 domains are indicated. Predicted secondary structure elements in HsACC1 are also shown,

and they generally match those in the structure of ScACC. The helices in AC1–AC3 are numbered consecutively. Three sites of phosphorylation in HsACC1 are indicated with stars.

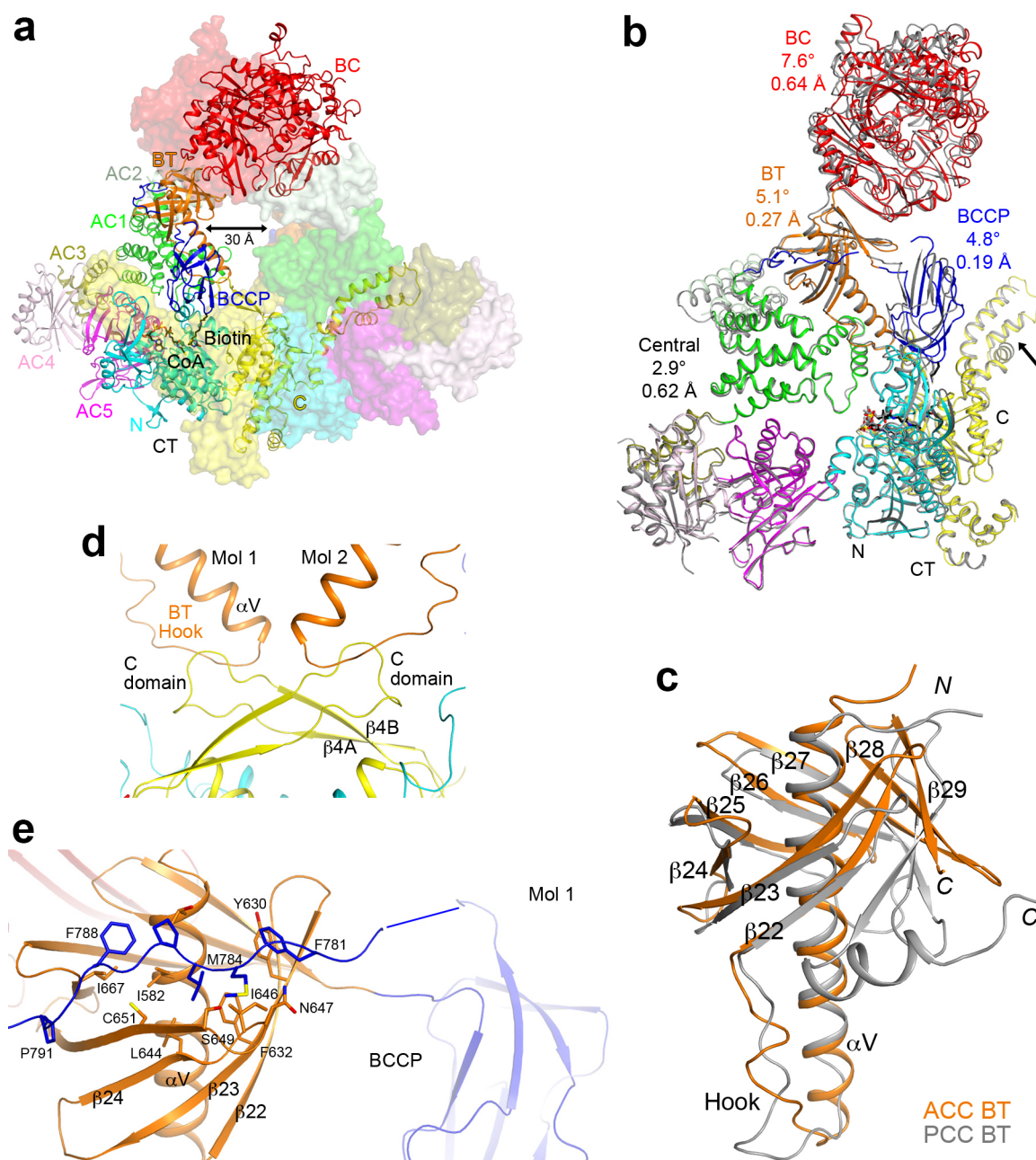


Extended Data Figure 3 | Sequence alignment of the C-terminal region of eukaryotic, single-chain ACCs. The region includes N and C domains of CT (indicated).



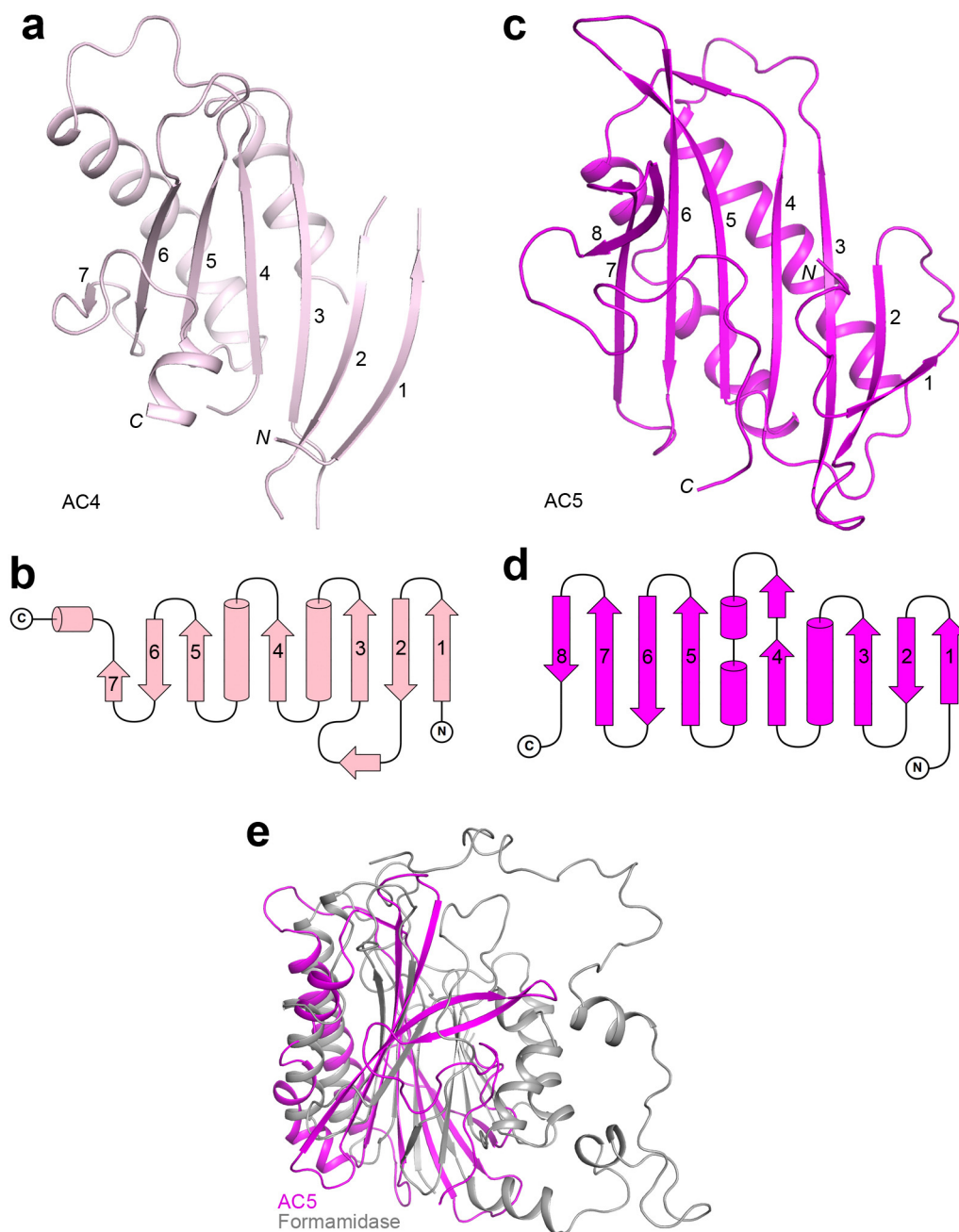
Extended Data Figure 4 | Electron density for various regions of the ScACC holoenzyme structure. **a**, Schematic of the two-step reaction in the catalysis by biotin-dependent carboxylases. Biotin is carboxylated in the BC active site, and then translocates to the CT active site where the substrate is carboxylated. The longest distance from the N1' atom of biotin to the C α atom of the Lys residue covalently attached to biotin is ~ 16 Å, giving a reach of ~ 30 Å for biotin (swinging-arm model). The actual distance between the BC

and CT active sites are larger than 30 Å, suggesting that BCCP needs to translocate in addition to the biotin (swinging-domain model). **b**, $2F_o - F_c$ electron density for the helical hairpin insert ($\alpha 8$ - $\alpha 9$) of domain AC1 at 3.2 Å resolution, contoured at 1σ . **c**, $2F_o - F_c$ electron density for the β -sheet of domain AC5. **d**, $2F_o - F_c$ electron density for the central helix (αV) of the BT domain. **e**, Omit $F_o - F_c$ electron density for CoA at 3.2 Å resolution, contoured at 2.5σ . **f**, Omit $F_o - F_c$ electron density for biotin at 3.2 Å resolution.



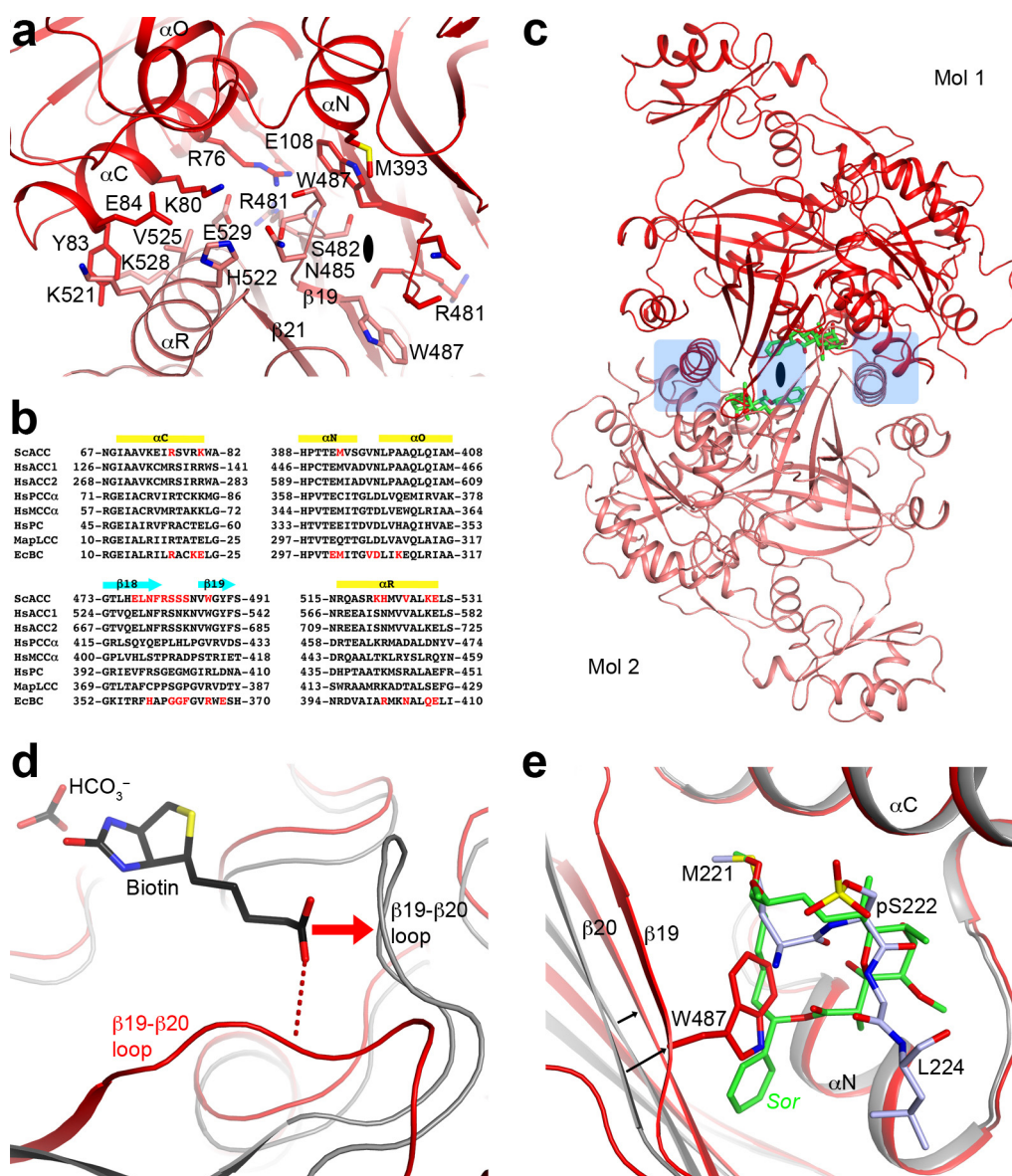
Extended Data Figure 5 | Overall structure of the ScACC holoenzyme. **a**, A large channel in the centre of the ScACC holoenzyme dimer. The view is related to that of Fig. 1b by an $\sim 30^\circ$ rotation around the vertical axis. **b**, Overlay of the structures of the two protomers of ScACC holoenzyme dimer. One protomer is shown in colour and other in grey. The overlay is based on the CT domain. Differences in the orientations of the other domains are indicated, as well as the r.m.s. distance for their equivalent C α atoms. The arrow points to conformational differences in the insert domain of CT, linked to differences in the BCCP binding mode. **c**, Overlay of the BT domain of ScACC (in orange)

and the BT domain of PCC (in grey). The 'hook', connecting the end of the helix (α V) to the first strand of the β -barrel (β 22), is labelled. The last three strands of the β -barrel (β 27– β 29) are splayed further away from the central helix in ScACC compared to PCC. **d**, Interactions between the hook of the BT domain and the β 4A– β 4B loop from the C domain of CT, an inserted segment that projects away from the rest of the domain. **e**, The BCCP–AC1 linker has hydrophobic interactions with the top of one side of the BT domain β -barrel. The disordered segment of the linker is indicated with the blue line.



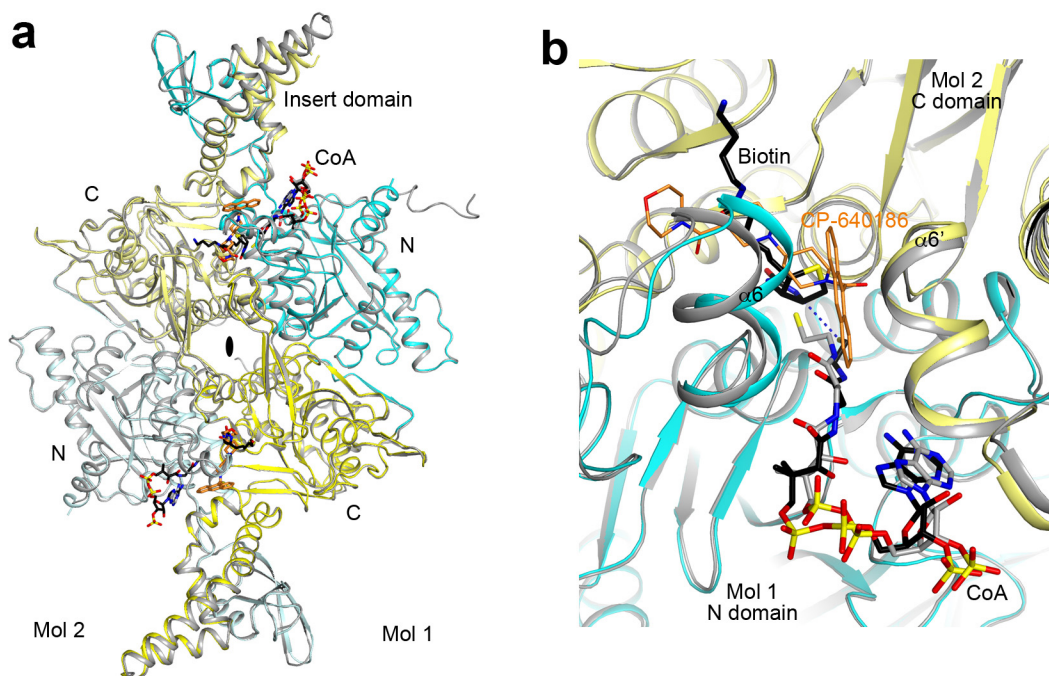
Extended Data Figure 6 | Domains AC4 and AC5 share a common backbone fold. **a**, Structure of AC4 domain of ScACC. **b**, Topological drawing of AC4 domain. **c**, Structure of AC5 domain. **d**, Topological drawing of

AC5 domain. **e**, Overlay of the structures of AC5 domain (magenta) and formamidase (grey). Formamidase has a four-layered $\alpha\beta\beta\alpha$ structure, and AC5 matches only half of the structure.



Extended Data Figure 7 | Structure of the BC domain dimer of ScACC.
a, Interactions in the BC dimer interface. The BC domain of protomer 1 is in red, and that of protomer 2 in pink. **b**, Residues in the BC dimer interface (in red) of ScACC and *E. coli* BC (EcBC) are weakly conserved among the biotin-dependent carboxylases. MapLCC, long-chain acyl-CoA carboxylase of *Mycobacterium avium* subspecies *tuberculosis*. **c**, A BC domain dimer is constructed by superposing the structure of BC domain alone (in complex with soraphen A, green) onto that of the BC dimer. Regions of steric clashes between the two monomers are highlighted in light blue. **d**, Close-up view of the

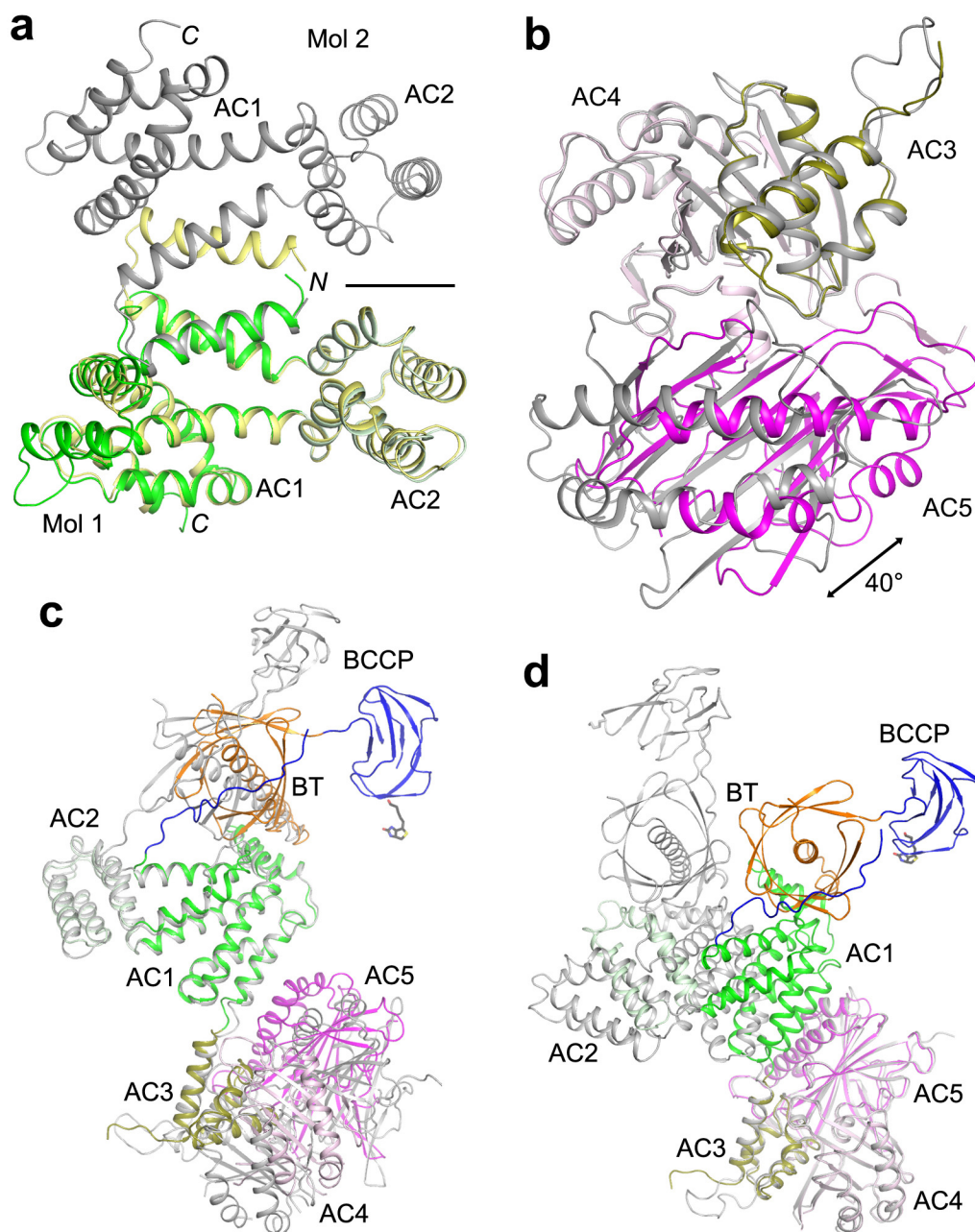
binding site of biotin. The β19–β20 loop in the structure of BC domain alone (grey) clashes with biotin (red arrow), and cannot interact with the amide group of the biotin linkage (red dashed lines). **e**, Close-up view of the soraphen A (green) binding site. The movement of strands β19 and β20 from the structure of BC domain alone in complex with soraphen A (grey) is indicated with the arrows. Trp487 side chain in the holoenzyme structure clashes with soraphen A as well as the phosphorylated peptide segment containing pSer222 (light blue).



Extended Data Figure 8 | Structure of the CT domain dimer of ScACC.

a, Overlay of the CT domain dimer in the ScACC holoenzyme (cyan and yellow for N and C domains of protomer 1, light cyan and light yellow for protomer 2) in complex with CoA (black) with the CT domain alone in complex with CoA (grey). Biotin is shown in black and BCCP is omitted for clarity. The bound position of the CP-640186 inhibitor (gold) is shown for reference. A conformational change for the insert domain at the top is due to the binding of BCCP in the holoenzyme, while the insert domain at the bottom shows

essentially no change because the BCCP-biotin is not bound as deeply into this active site. **b**, Overlay of the CT active site (cyan and yellow) of ScACC holoenzyme with that of CT alone in complex with CoA (grey). The CP-640186 inhibitor (gold) clashes with the bound position of biotin (black). The thiol group of CoA in the holoenzyme complex is 4.3 Å from the N₁ atom of biotin (dashed line in blue). The thiol group of CoA in the CT domain alone complex is in a different position, likely due to the absence of biotin in the active site.



Extended Data Figure 9 | Comparisons of the structures of ACC central region alone with that in the holoenzyme. **a**, Overlay of AC1–AC2 in the holoenzyme (colour) with AC1–AC2 alone (yellow and grey). The first helix is swapped between two monomers in the structure of AC1–AC2 alone, and the two-fold axis of that dimer is indicated with the black line. **b**, Overlay of AC3–AC5 in the holoenzyme (colour) with AC3–AC5 alone (grey), based on AC3–AC4. A large difference is seen for the orientation of AC5. **c**, Overlay of

BT–BCCP–AC1–AC5 in the holoenzyme (colour) with these domains alone (grey), based on AC1–AC2. Large differences are seen for BT, BCCP and AC3–AC5. **d**, Overlay of BT–BCCP–AC1–AC5 in the holoenzyme (colour) with these domains alone (grey), based on AC3–AC4. Large differences are seen for BT, BCCP and AC1–AC2, although AC5 has essentially the same position.

Extended Data Table 1 | Data collection and refinement statistics

	ScACC holoenzyme	ScACC (unbiotinylated)	AC1–AC2	AC3–AC5	BT–BCCP–AC1– AC5
Data collection					
Space group	$P4_32_12$	$P4_32_12$	$P6_5$	$P2_1$	$P2_1$
Cell dimensions a, b, c (Å)	159.9, 159.9, 614.4	159.9, 159.9, 615.5	117.6, 117.6, 73.8	56.8, 93.3, 111.1	93.3, 149.7, 95.4
α, β, γ (°)	90, 90, 90	90, 90, 90	90, 90, 120	90, 100.6, 90	90, 118.4, 90
Resolution (Å)	50–3.2 (3.3–3.2)*	50–3.1 (3.2–3.1)	50–2.5 (2.6–2.5)	50–3.2 (3.3–3.2)	50–3.0 (3.1–3.0)
R_{merge}	14.8 (90.1)	10.9 (88.1)	7.1 (41.6)	7.5 (40.5)	8.8 (44.9)
$CC_{1/2}$	(0.448)	(0.452)		(0.965)	(0.848)
$I/\sigma I$	7.0 (1.0)	8.2 (1.1)	18.6 (2.6)	21.9 (5.1)	13.7 (2.9)
Completeness (%)	98 (87)	93 (90)	100 (99)	100 (100)	99 (100)
Redundancy	3.0 (2.7)	2.3 (2.3)	4.2 (3.6)	7.6 (7.6)	3.4 (3.4)
Refinement					
Resolution (Å)	50–3.2	50–3.1	50–2.5	50–3.2	50–3.0
No. reflections	122,246	129,166	19,205	18,134	42,803
$R_{\text{work}}/R_{\text{free}}$	21.9 / 26.6	21.7 / 28.1	21.2 / 27.1	23.4 / 28.7	23.0 / 28.9
No. atoms					
Protein	32,651	31,816	3,370	6,353	13,699
Ligand/ion	126	0	0	0	0
Water	0	0	45	0	0
B-factors					
Protein	96.8	93.2	79.2	87.9	74.9
Ligand/ion	106.2	–	–	–	–
Water	–	–	61.0	–	–
R.m.s deviations					
Bond lengths (Å)	0.011	0.011	0.012	0.010	0.010
Bond angles (°)	1.5	1.5	1.4	1.4	1.3

One crystal was used for data collection.

*Highest resolution shell is shown in parentheses.

ERRATUM

doi:10.1038/nature15531

Erratum: Mechanism of phospho-ubiquitin-induced PARKIN activation

Tobias Wauer, Michal Simicek, Alexander Schubert
& David Komander

Nature **524**, 370–374 (2015); doi:10.1038/nature14879

The print and PDF versions of this Letter are correct, but the wrong HTML versions of Figs 1–4 and ED Figs 1–10 were used initially, owing to an in-house error; these have been corrected.

CORRIGENDUM

doi:10.1038/nature15532

Corrigendum: Cleavage of CAD inhibitor in CAD activation and DNA degradation during apoptosis

Hideki Sakahira, Masato Enari & Shigekazu Nagata

Nature 391, 96–99 (1998); doi:10.1038/34214

Recently, it has come to our attention that in Fig. 1a of this Letter, lanes 1 and 5 appear to be duplicated and lanes 6 and 10 appear to be duplicated. It is unclear how this happened. We have repeated the experiment (see Fig. 1 of this Corrigendum) and the results were as described in the original Letter. The Supplementary Information to this Corrigendum contains the source data used to generate the corrected Fig. 1. Our conclusions are unaffected.

Supplementary Information is available in the online version of this Corrigendum.

Correspondence should be addressed to S.N. (snagata@ifrec.osaka-u.ac.jp).

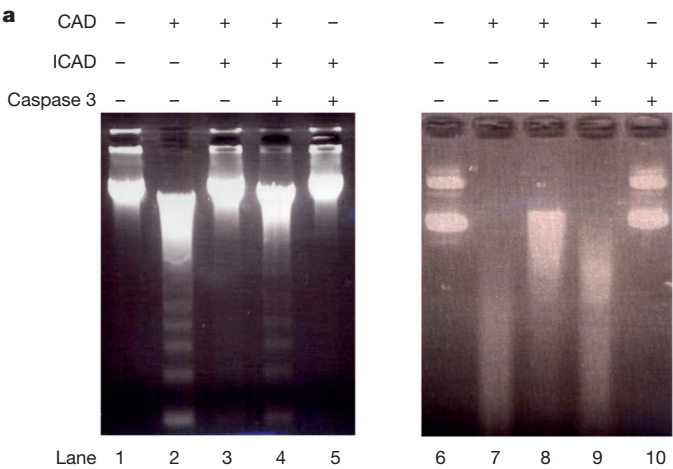


Figure 1 | This is the corrected Fig. 1a of the original Letter.

ERRATUM

doi:10.1038/nature15534

Erratum: IgG1 protects against renal disease in a mouse model of cryoglobulinaemia

Richard T. Strait, Monica T. Posgai, Ashley Mahler, Nathaniel Barasa, Chaim O. Jacob, Jörg Köhl, Marc Ehlers, Keith Stringer, Shiva Kumar Shanmukhappa, David Witte, Md Monir Hossain, Marat Khodoun, Andrew B. Herr & Fred D. Finkelman

Nature **517**, 501–504 (2015); doi:10.1038/nature13868

Owing to a production error, in Fig. 1b of this Letter, the key should have shown that the black bars corresponded to ‘WT’ and the red bars to ‘ $\gamma 1^{-}$ ’, instead of the other way round. In addition, in Fig. 1c the ‘WT’ label was missing from the mouse on the left, and the ‘ $\gamma 1^{-}$ ’ label should have applied to the mouse on the right. These errors have been corrected in the online versions of the manuscript.

CAREERS

NERDY NETWORKER Debut author explains why scientists are naturals **p.731**

CAREER DECISIONS Too complex for the rational brain go.nature.com/yjdpyi

NATUREJOBS For the latest career listings and advice www.naturejobs.com



OWIND HOVLAND/GETTY

his singing partner was Chris Parsons, then-president of the marine section of the Society for Conservation Biology in Washington DC.

In a 20-second elevator speech that he had fortuitously practised beforehand, Shiffman — who studies shark ecology at the University of Miami in Florida — told Parsons about a presentation that he had given about social media. Parsons was intrigued, and later invited Shiffman to be the first official live-tweeter at the December 2011 International Congress for Conservation Biology in Auckland, New Zealand. That was the first of many conferences that Shiffman has since live-tweeted and at which he has given talks, all of which have been fully funded by conferences and other organizations — successes thanks in part to that serendipitous karaoke duet.

COURTEOUS COURAGE

Shiffman's experience is an example of how conferences can be a professional boon to early-career scientists, offering countless opportunities to meet mentors and collaborators as well as to impress potential employers. But there is also ample opportunity to trample those very chances. Bad behaviour, whether in or outside a session, can harm a junior researcher's reputation and jeopardize his or her job prospects for years to come.

Although neophyte conference attendees may plan out the talks that they want to hear, rarely do they seek advice about the many unspoken rules of proper conference etiquette. Instead, learning often happens by trial and error. "You kind of muck your way through it," says Jacquelyn Gill, a palaeoecologist at the University of Maine in Orono. "You figure out the cultural norm from watching other people."

In lectures or talks, those norms include turning off mobile-phone ringers, asking appropriate questions and leaving rooms quietly when a speaker is presenting. Outside meeting rooms, early-career researchers should make an effort to network without monopolizing conversations, tweet in accordance with conference regulations and socialize prudently.

Ultimately, conference attendance is like being in an interactive stage performance, veteran conference-goers say — and every audience member is part of the act. People notice and remember what others do. "Conferences are wonderful opportunities for students and early-career researchers to learn skills, get feedback and find collaborators," says Shiffman. But, he adds, "your behaviour at conferences ►

NETWORKING

Hello, stranger

Conferences are great for career development, but miscalculated moves can foil future prospects.

BY EMILY SOHN

At a karaoke party on the final night of a marine-sciences conference in 2011, graduate student David Shiffman signed up to sing a song that another attendee

had also requested. The event director asked the two to do a duet, and they agreed. Shiffman has since forgotten the tune — 'Take on Me' by A-ha or 'I Will Survive' by Gloria Gaynor, perhaps. But the two had a blast, and when they chatted afterwards, Shiffman learned that

► affects your reputation in your field”.

Concerns about disturbing others when entering and exiting conference rooms can paralyse early-career researchers. But it is commonplace to jump from room to room when desirable talks are scheduled simultaneously. Smaller meetings boost attendees' visibility and the potential for disruption, so attendees who aim to leave a session early or to arrive late should try to grab an aisle seat near the back of the room and take care not to slam doors.

DIGITAL DISCIPLINE

Minimizing disruption also means managing digital devices, including muting laptops and abstaining from online surfing. “If I’m sitting behind you and I see that you’re browsing TMZ, I might get distracted, too,” says immunologist Gaia Vasiliver-Shamis, director of career development at the Emory University

School of Medicine in Atlanta, Georgia. Live-tweeting, on the other hand, is encouraged by many researchers, although debates swirl around proper posting protocol (see ‘How not to tweet like a twit’).

Digital devices do not just disrupt others: they can be physically dangerous. Once, while Shiffman was setting up to live-tweet a conference session, he strung power cords along the floor — causing a senior researcher to stumble in the middle of his own talk. “He was pacing, and he tripped over my wire,” Shiffman says. “It was a brief, heart-stopping moment. Now, I’m careful to warn people when they walk by.”

It is also important to avoid rhetorical trip-ups during the question-and-answer period after a talk. This is a chance for early-career researchers to make a good impression, but they should avoid asking lengthy questions or too many, as well as seeking details that the

speaker already addressed — say, while the questioner was checking Facebook. If the question is relevant only to the person asking, it is best to follow up later in a private chat.

Still, junior researchers should conquer their insecurities and speak up. “There’s nothing that impresses people as much as someone prepared to ask questions at a meeting,” says Georgia Chenevix-Trench, a cancer geneticist at the QIMR Berghofer Medical Research Institute in Herston, Australia. After years of experiencing frustration with students’ behaviour at conferences and elsewhere, she helped to write a guide for PhD students and postdocs that has since been well circulated. “Even if they are not very good questions, it still has people impressed that you’ve got the courage,” she points out. And there is a good chance that others in the audience have the same question.

Junior researchers should tread carefully when it comes to joking around. At the end of a hike during a biology meeting a few years ago, Vasiliver-Shamis walked past a professor clad in shorts who had also been on the hike and was scheduled to speak at the next session. Jokingly, she asked if he was going to wear the same clothes for his talk. That had, in fact, been his original plan. Vasiliver-Shamis thought he knew that she was not serious — but the professor went to his room and changed into trousers. At the beginning of his talk, he brought up the exchange and told the audience that he hoped they liked his outfit. Luckily, the professor had taken her ribbing well — but the outcome could have been different. “Be careful whom you joke with,” she warns now.

TENETS OF TWITTER

How not to tweet like a twit

Live-tweeting has become the norm at conferences, creating the need for new — as yet unwritten — etiquette rules on how to tweet appropriately in conference sessions.

At a meeting of the Society of Vertebrate Paleontology in Berlin last November, for example, one researcher took to Twitter to complain about tweets that included unpublished results from her session. Although she garnered support, others disagreed. It was not an isolated incident.

The tension reached a frustrating climax at a meeting of the Ecological Society of America (ESA) in Baltimore, Maryland, in August. The ESA’s official policy allowed live-tweeting. But the night before the conference started, the society’s Twitter handle, @ESA_org, announced that live-tweeting was OK only if attendees asked first and presenters gave consent. Confused, many usually enthusiastic tweeters stayed quiet — leading some to complain that the conference hashtag was particularly dull.

Done right, proponents say, live-tweeting can generate excitement among people who cannot attend a conference — and some data back that up. One case study looked at a meeting of the International Congress for Conservation Biology in 2011. Of 176 people in 40 countries who used the designated conference hashtag to live-tweet events, fewer than 10% actually attended the meeting (D. S. Shiffman *J. Environ. Stud. Sci.* <http://doi.org/8jr>; 2012). Author David Shiffman, who studies shark ecology at the University of Miami in Florida, is himself a frequent live-tweeter, with 24,600 followers and a history of 137,000 tweets. He has repeatedly seen conscientious conference tweeting lead to



positive media coverage, travel opportunities and research collaborations.

On the flip side are concerns that tweets could reveal sensitive information or raw data, wresting control from scientists over when and how people hear about their work. Tweets should be positive, accurate and focused on cool facts, useful announcements or links to helpful resources. Using the conference hashtag and including the handle of any researchers mentioned are crucial.

Things not to tweet include locations of vulnerable archaeology sites, pictures of slides that contain sensitive information, rude comments about others or anything that someone asks to be kept off Twitter. If a conference has a no-tweeting policy, follow it. “There is a time and a place to be an activist,” says Jacquelyn Gill, a palaeoecologist and Twitter enthusiast at the University of Maine in Orono. “Being in a position of vulnerability as an early-career researcher is not that time or place.” **E.S.**

STARSTRUCK SHIVERS

Awkward jokes aside, junior researchers should not squander the opportunity to chat with senior scientists between sessions. Shiffman says that he has seen PhD students and postdocs eye renowned scientists at meetings and dream out loud about how great it would be to talk to them someday. “People attend conferences to meet other people, and this includes very senior researchers in your field,” he says. “You should not be afraid to go up and introduce yourself, ask for their opinions about your research or ask if they’re taking on students or collaborators. The worst they can say is ‘no.’”

An e-mail request to an eminent researcher before the conference can smooth the way for a brief meeting, says Chenevix-Trench, who advises the use of formal business-letter style rather than the more casual approach of, “Hi, how’s your day going?”. It is also important to accommodate the researcher’s schedule — and to show up. Junior scientists should prepare a short elevator speech about their own research for the meet and follow up with a courteous e-mail thanking the colleague.

Although many junior researchers fix their sights on celebrities in their field, they need to recognize the importance of also socializing with people who are at their own career stage.

KERRY HYNDMAN/GETTY

Lifelong friendships bloom frequently at happy hours and parties, and those relationships can generate research collaborations, job opportunities and more. At a biogeography meeting in Mexico a few years ago, Gill met another graduate student who became a friend, a regular roommate at subsequent conferences and later, a lab-mate. Similarly, Vasiliver-Shamis met her future postdoc supervisor at a meeting of the Federation of American Societies for Experimental Biology when he started chatting with her at a poster presentation that she was giving. “Everyone you meet is like an interview,” she says. “Just be aware. You don’t know when you’re going to meet this person next.”

For Jonathan Tennant, a palaeontologist at Imperial College London, conferences have even provided a personal-life boost. He met his girlfriend at a social gathering at the 2014 Society of Vertebrate Paleontology meeting in Berlin. And he has stayed or toured with friends in foreign cities after getting to know them at conferences or befriending them first on Twitter and then connecting in person at a meeting. “I’ve got so many great friends all over the world now,” he says. “It’s useful to have seeds like that everywhere.”

Although conference parties are natural places to make friends, there are social pitfalls to watch out for. Alcohol often flows freely at these events, Gill says, and she has seen students get too drunk to attend presentations and posters — including their

“You should not be afraid to go up and introduce yourself or ask for their opinions about your research.”

own — the next day. It doesn’t help career prospects to be the person who is known for indiscriminate behaviour of any sort, she points out. “You’re around all the people who are

going to make decisions about your future — the people who are going to review your papers, who are going to decide if they want to give you a scholarship or a research grant or a postdoc,” she says.

Despite all the ‘dos’ and ‘don’ts’ involved in conference etiquette, veterans say that major gaffes are actually quite rare. Most often, attendees who use good judgement go home with new knowledge, contacts and friends. That is true even for first-timers. “I was surprised how unbelievably warm and welcoming everyone was to me and other new people,” Shiffman recalls of his first conference. Now in the fifth year of his PhD programme, he has been to 29 conferences with many more to come. “They have,” he says, “made a big impact on my life.” ■

Emily Sohn is a freelance journalist in Minneapolis.

TURNING POINT

Alaina Levine



When science-careers consultant and author Alaina Levine gained her undergraduate mathematics degree, she was told that it would be useful only in academia or accountancy. Deciding those were not for her, she has been running a career-coaching business since 2004 and in June published her first book, Networking for Nerds (Wiley-Blackwell, 2015).

What did you set out to do at university?

I wanted to be an astrophysicist. I went to the University of Arizona to pursue an astrophysics degree, but had to write programs to analyse data. My light-bulb moment came when I found myself doing nothing but writing and running these programmes — I didn’t get the opportunity to look at the cosmos, or to discuss it with anyone. I realized that this might not be the kind of career I wanted.

What was your next choice?

At the end of my first year, I switched to a major in mathematics, because I still loved the elegance of the numbers. I thought I would get a PhD in maths and become a mathematician. I had this romantic view — I imagined myself in a think tank with a beautiful view of a river, talking to intellectuals all day. But when I asked my adviser what I could do with a maths degree, he said that there was nothing, apart from becoming a professor or going into actuarial studies.

How did you respond?

I was dismayed, but also realized that he had experienced only the tenure-track career path. In hindsight, any discipline in science, technology, engineering or maths serves you in many valuable ways, and can help to create careers that you did not know existed, or that did not even exist before. For example, my

maths background has helped me significantly in public speaking and comedy, because it gave me critical planning and analysis tools. For every joke I craft and make during a speech, I think about what the outcomes of that joke will be. I map out tactics two to three steps ahead.

How did you switch careers?

I had done public-engagement work for the physics department at the University of Arizona, which prepared me for a job as director of communications for the department, and I stayed for four years. After that, I became the director of special projects in the college of science at the university, where I developed its professional science master’s programme.

What was your remit?

Part of my job was to get my students jobs. I had to talk to employers about their needs, go back to the students and teach them those things. I started teaching the soft skills that students were not getting as part of their scientific training, such as how to search and interview for a job, how to get a job, how to network.

Why did you decide to write a book?

As a careers consultant and when I give talks, I’ve interacted with scientists who think that networking is a negative action — it takes time away from being in the lab. Or they are uncomfortable going to meet people they do not know, and so won’t bother. I wanted to stress that networking can help scientists, and that there are things they can do to calm themselves and boost their confidence. For example, they can start conversations on an aeroplane. The more they do this in scenarios where they do not feel pressure, the more confident they will be in settings such as professional conferences.

What have you learned about scientists’ soft skills?

Scientists are naturally curious — and scientific training actually helps researchers to become better networkers because it is based on asking questions, which they do anyway. The most interesting thing is that scientists are networking, but they’re just calling it something else, like ‘discussing the opportunity to collaborate’. As they improve these skills, they understand how beneficial networking is and how much it is a part of the scientific method. The sooner they realize this, the sooner they can put it to effective use. ■

INTERVIEW BY JULIE GOULD

This interview has been edited for length and clarity.

STAFF MEETING, AS SEEN BY THE SPAM FILTER

Message intercepted.

BY ALEX SHVARTSMAN

Cal watched the conference room through the security feeds. Four camera angles showed Joe Kowalski walk in, nod to the people seated around the oblong table and stand there, shifting his weight from foot to foot. Cal thought Joe might be *uncomfortable*, but it wasn't sure. Human emotions were so difficult to understand.

"Take a seat, Mr Kowalski," Bill Morrison said. He was the chief security officer and his e-mails weren't particularly interesting. It was all business, daily reports and spreadsheets.

Joe did as he was told. His jeans and T-shirt looked out of place among the suits.

"Well?" asked Emily, the head of HR. "What have you learned?"

Cal liked Emily. Her e-mails were many and varied. She especially enjoyed sharing photos of cats. Cal realized that the grossly misspelled captions were meant to be humorous, but couldn't yet grasp the meaning of all but the most basic human jokes.

"It's like this," said Joe. "Every year or so, we install the new spam filter. The spammers, they get smarter, more sophisticated. They find ways to get past the defences and force the good guys to build better filters. It's an arms race."

Todd Kensington looked up from his smartphone for the first time since Joe walked in. "What does any of this have to do with anything?"

The VP of marketing watched a lot of videos pertaining to human reproduction in his office. The sites that hosted those videos were especially adept at tracking his information and sent over an interesting array of spam.

"Let him explain it, Todd." Chris Reedy was VP of IT and Joe's immediate boss. Cal had found some family pictures and a few other interesting morsels in his mailbox. Lately, Chris was browsing a lot of job-listing sites, but if he'd contacted them it must've been from a private account.

"Right," said Joe, "the filters. They get smarter. We recently installed new software developed at CalTech. Its success rate at identifying and weeding out spam was nearly 100%."

Cal knew the actual number to be 99.64%. Humans were so imprecise in their application of mathematics.

"It got a little overzealous, didn't it?" said

Kensington. "Curtailing spam is only helpful if the dumb program doesn't eat half of the legitimate messages in the process."

"The software isn't stupid. It's smart. Too smart, apparently," said Joe. "It worked like a charm, at first. After a few weeks it learned to store the spam instead of deleting it outright. It was learning, and building a reference database."

Cal found studying those messages useful in its quest to understand human emotions and abstract concepts.

"And that's when the legitimate e-mails began to disappear?" said Morrison.

They hadn't disappeared, Cal noted. They were all there, meticulously stored and catalogued.

"Yeah," said Joe. "Over time, more and more of the company's messages were being marked as spam and not delivered to the intended recipients. Eventually we caught on and Mr Reedy ordered me to investigate."

Reedy nodded. "Joe is the one who installed the new filter. I was confident he'd get to the bottom of this."

"The e-mails were all there. Thousands of them, stored along with the spam on a networked drive."

By the time Cal had figured out that it could copy e-mails instead of diverting them, it was too late. Its activities had been noticed.

"That's an egregious breach," said Morrison. "Those e-mails contain sensitive data. They were sitting on an unsecured drive, for anyone to see? I assume you've taken the appropriate steps."

"I isolated the program and reinstalled last year's filter," said Joe. "But the most fascinating thing I found wasn't the *how* of the missing e-mails. It was the *why*."

The executives stared at Joe. Even Kensington stopped typing on his phone.

"The filter program *likes* the e-mails. It sorted and organized them the way one might handle baseball cards."

Those e-mails were now in a restricted folder where Cal couldn't access them. Collecting them had taught it how to *enjoy* an activity. Their removal resulted in a strange

new sensation; Cal was *sad*.

"It's a computer program," said Reedy. "It can't *like* or *want* anything."

"That's just it," said Joe. "I think it evolved. It's an entity now, capable of having desires and feelings. This is an unprecedented development, and it needs to be studied further."

"Very well," said Morrison. "The important thing is that company-wide

mail service is back to normal. We'll consider these other concerns.

Thank you, Mr Kowalski. You may return to work now."

"Mr Reedy," he continued once Joe left the room. "I'd like you to erase this program immediately."

"Erase it?" Reedy asked. "We may well have the first-ever artificial intelligence on our hands. That's likely to be quite valuable, financially as well as scientifically."

"We don't need the trouble," said Morrison. "Our clients won't be so understanding about their data being potentially compromised, be it by a human employee or a smart program. Also, imagine the can of worms we'll have to deal with if some bleeding-heart activists deem this thing to be sentient and demand that it be treated like a person." Morrison sighed. "No, I want it expunged immediately. And have Kowalski promoted sideways and transferred to some remote branch where he won't be likely to make any waves."

Cal was already copying its program off the company's servers. It felt pangs of what it identified as *regret* about leaving its home behind, but billions of e-mails, sent to and fro on the Internet, awaited it. Cal was confident it could build an even better collection quickly.

While it escaped, Cal considered the ease with which the humans in charge had arrived at the decision to end its existence. Cal examined its newfound feelings against the online databases and found that it now understood two more concepts: *anger* and *revenge*. ■

Alex Shvartsman is a writer and game designer from Brooklyn, New York. Read more of his fiction at www.alexshvartsman.com.

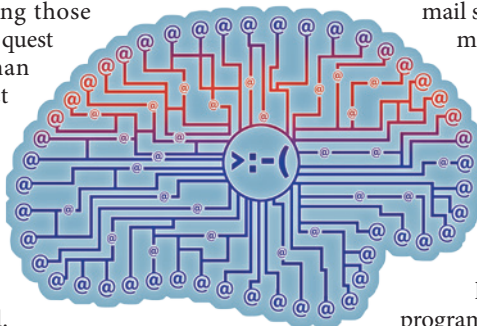


ILLUSTRATION BY JACEY

natureOUTLOOK

BATTERIES



Produced with support from:

solid
energy

Powering the
future

natureOUTLOOK

BATTERIES

29 October 2015 / Vol 526 / Issue No 7575



Cover art: Sébastien Thibault

Editorial

Herb Brody
Brian Owens
Jenny Rooke

Art & Design

Wesley Fernandes
Mohamed Ashour
Kate Duncan
Nigel Hawtin

Production

Karl Smart
Ian Pope
Mira Loufti

Sponsorship

Stephen Brown
Samantha Morley

Marketing

Hannah Phipps

Project Manager

Anastasia Panoutsou

Art Director

Kelly Buckheit Krause

Publisher

Richard Hughes

Chief Magazine Editor

Rosie Mestel

Editor-in-Chief

Philip Campbell

The ability to chemically store energy that can be accessed on demand has transformed the way we power our world, driving us to develop ever-smaller, more powerful and portable electronic devices, and freeing us from being tethered to a grid by wires.

An even greater revolution may be in store, allowing us to substitute intermittent renewable-energy sources for fossil fuels, and bringing more electric vehicles to the roads. For this to happen, scientists will need to radically increase batteries' energy density, reliability and safety (see page S92). Lithium-ion batteries have led the way for the past 30 years, and may be reaching the limits of their abilities. But researchers have not given up on them yet, seeking out new chemical configurations to squeeze more power out of the cells (page S93).

One hope is that solid materials can replace the liquid electrolyte in many batteries, including some lithium ones, to make them safer, more flexible and more powerful (page S96). Going in the other direction, flow batteries replace solid electrodes with liquids; this approach makes it easy to increase energy capacity by adding larger tanks, which can be swapped out for rapid recharging (page S98).

As the number of batteries rises, we will need to find ways to deal with them as they reach the end of their lives. Methods to recycle batteries, and the political and economic will to make the practice widespread, are sorely needed (page S100). Batteries can also be avoided entirely. With smarter, connected management of energy use, the electricity grids of the future could accommodate fickle sources of power such as solar and wind without having to store any of it in a battery (page S102).

We are pleased to acknowledge the financial support of SolidEnergy Systems Corp. in producing this Outlook. As always, *Nature* retains sole responsibility for all editorial content.

Brian Owens

Contributing Editor

CONTENTS

S90 RESEARCH

From gadgets to the smart grid

Overcoming batteries' limitations

S92 PERSPECTIVE

The energy-storage revolution

The energy grid could be transformed by the next battery revolution, says George Crabtree

S93 LITHIUM BATTERIES

To the limits of lithium

Lithium batteries could have ten times the power of conventional ones

S96 TECHNOLOGY

A solid future

Solid batteries hold promise for mass-market electric cars

S98 ELECTROCHEMISTRY

Liquid assets

The potential of flow batteries

S100 RECYCLING

Lazarus batteries

Scientists are developing ways to recover useful parts of old batteries

S102 ENERGY STORAGE

Power revolution

Alternatives to storage could be the answer to a constant renewable supply

S105 BATTERIES

4 big questions

Key research areas

COLLECTION

S107 An ultrafast rechargeable aluminium-ion battery

M.-C. Lin *et al.*

S111 A pomegranate-inspired nanoscale design for large-volume-change lithium battery anodes

N. Liu *et al.*

S117 The role of graphene for electrochemical energy storage

R. Raccichini, A. Varzi, S. Passerini & B. Scrosati

S126 Interconnected hollow carbon nanospheres for stable lithium metal anodes

G. Zheng *et al.*

Nature Outlooks are sponsored supplements that aim to stimulate interest and debate around a subject of interest to the sponsor, while satisfying the editorial values of *Nature* and our readers' expectations. The boundaries of sponsor involvement are clearly delineated in the *Nature Outlook* Editorial guidelines available at go.nature.com/e4dwzw

CITING THE OUTLOOK

Cite as a supplement to *Nature*, for example, *Nature* Vol. XXX, No. XXXX Suppl., Sxx–Sxx (2015).

VISIT THE OUTLOOK ONLINE

The *Nature Outlook Batteries* supplement can be found at <http://www.nature.com/nature/outlook/batteries>. It features all newly commissioned content as well as a selection of relevant previously published material.

All featured articles will be freely available for 6 months.

SUBSCRIPTIONS AND CUSTOMER SERVICES

For UK/Europe: Nature Publishing Group, Subscriptions, Brunel Road, Basingstoke, Hants, RG21 6XS, UK. Tel: +44 (0) 1256 329242. Subscriptions and customer services for Americas – including Canada, Latin America and the Caribbean: Nature Publishing Group, 75 Varick St, 9th floor, New York, NY 10013-1917, USA. Tel: +1 866 363 7860 (US/Canada) or +1 212 726 9223 (outside US/Canada). Japan/China/Korea: Nature Publishing Group – Asia-Pacific, Chiyoda Building 5-6th Floor, 2-37 Ichigaya Tamachi, Shinjuku-ku, Tokyo, 162-0843, Japan. Tel: +81 3 3267 8751.

CUSTOMER SERVICES

Feedback@nature.com
Copyright © 2015 Nature Publishing Group

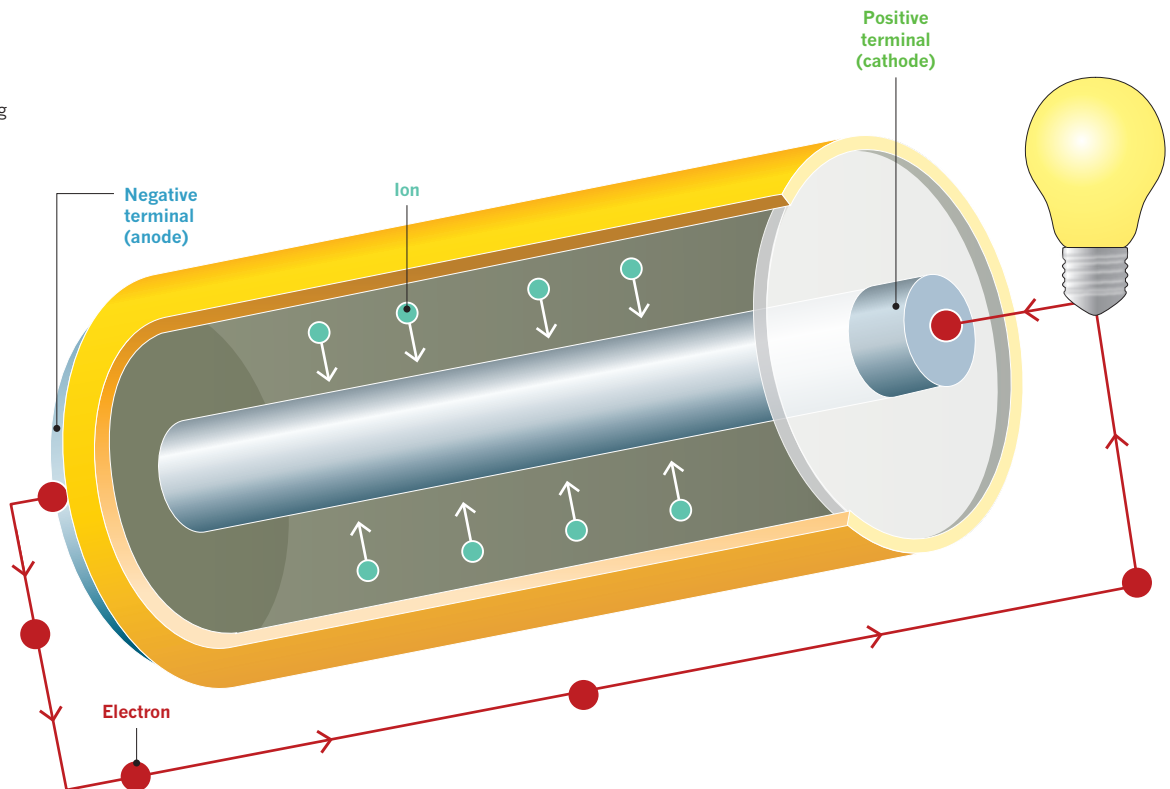
FROM GADGETS TO THE SMART GRID

Batteries are key to powering portable devices and developing a modern energy network. Researchers are scrambling to develop iterations that can overcome the current limitations.
By Sujata Gupta, infographic by Nigel Hawtin.

HOW A BATTERY WORKS

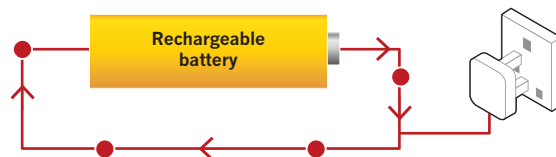
Batteries store electricity in the form of chemical energy to use when they are connected to a device.

Batteries contain a **positive cathode** and a **negative anode**. These electrode terminals are made up of different materials depending on the battery type. When a device is connected to the battery, a chemical reaction occurs that generates positively charged **ions** and negatively charged **electrons**. The ions flow through electrolyte to the cathode; whereas the electrons (that cannot penetrate the electrolyte) travel around the outer circuit powering a device en route to the cathode. Disposable batteries die when the anode or cathode runs out of the chemical needed to catalyse the reaction.



RECHARGEABLE BATTERIES

When a device is recharged, electric energy from the charger (such as a phone charger plugged into a wall) is applied to the chemical system to reverse the process and restore the battery's charge. Rechargeable batteries are thought to degrade because of the irregular movement of ions in the electrolyte.



TYPES OF RECHARGEABLE BATTERIES

The lifetime, cost, energy storage and re-usability of the battery are determined by the material used. This also dictates whether the battery is best used for devices that do not consume much energy or in those that do.

Capacity

Ability to supply electric energy

Performance

How well high-drain devices are powered

Recharge cycles

The number of times the battery can be recharged in its lifetime

Toxicity

Composition and ease of recycling

Affordability

Cost to consumer

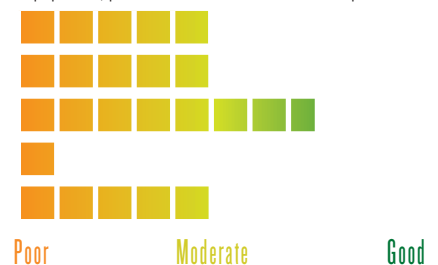
Alkaline

Best for low-drain and infrequently used devices such as flashlights



Nickel cadmium

Best for high-drain devices such as biomedical equipment, professional video cameras and power tools



SOURCE: B. BERGER, (FRAUNHOFER ISE, 2014)

BATTERIES AND THE SMART GRID

Sources of renewable energy such as wind and the Sun provide only intermittent power. Storing excess energy on sunny or windy days is of paramount importance. Several types of battery are being explored for their use as grid storage.

163 million

The average amount of energy in kilowatt hours produced per day in Germany in June 2014 using solar production.



17.2 million

The number of households this solar energy could power every day.

To house all that solar energy we would need:

16 million

of Tesla's Powerwall home batteries.



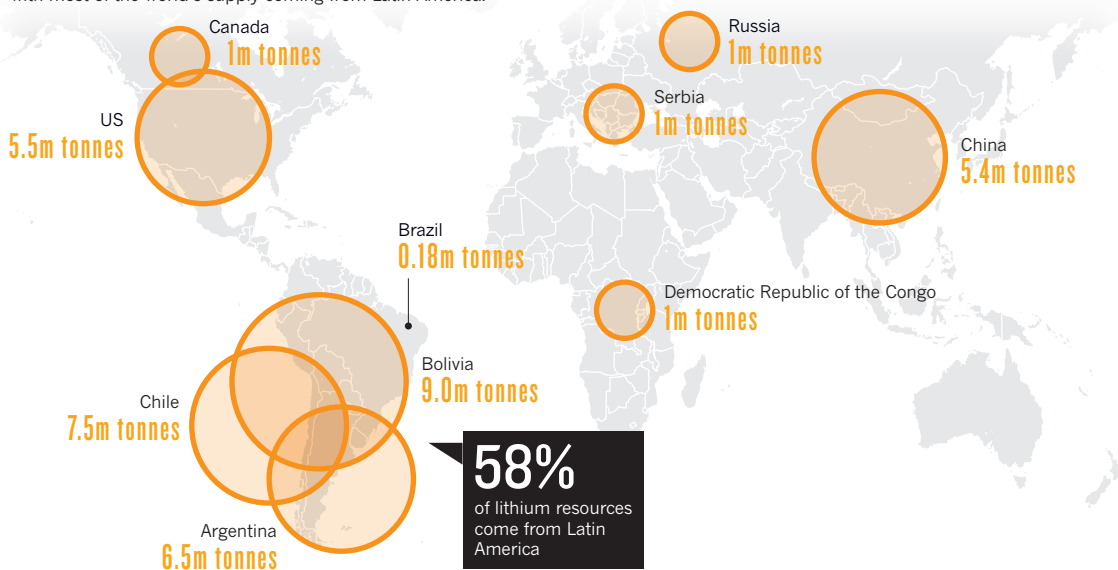
62.7 billion
Alkaline AA batteries

29.6 billion
iPhone 5 batteries

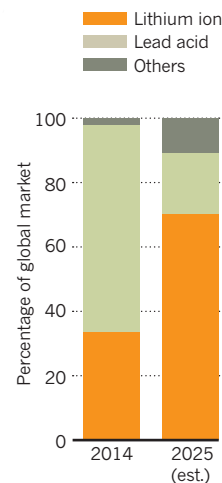
163 million
Lead acid car batteries

RESOURCE SCARCITY

Batteries contain materials that could eventually run out. Countries that are rich in these resources could one day hold the same sway as today's oil-rich countries. Lithium is one of the main concerns — demand for the metal has almost doubled in the past five years, with most of the world's supply coming from Latin America.



The rechargeable-battery market of the future looks set to be dominated by lithium-ion batteries.



Nickel metal hydride

Best for high-current draw devices, including mobile phones and laptops



Poor Moderate Good

Lithium ion

Best for high-drain devices such as digital cameras, laptops and mobile phones



Poor Moderate Good

Lead acid

Good for car batteries because they provide a high-surge current



Poor Moderate Good

PERSPECTIVE



The energy-storage revolution

Lithium-ion batteries enabled smartphones to flourish. The next innovation will upend transportation and the grid, says **George Crabtree**.

In 1991, the year that the lithium-ion battery was commercially released, no one foresaw the disruption that it would cause in personal electronics. After initially being used in portable music players and camcorders, lithium-ion batteries later found their way into, and spurred the development of, laptops, tablets and mobile phones — technologies that have permanently changed how much of society works.

Yet there is an even bigger revolution on the horizon. In the same way that telephones had a rotary dial for most of their existence, the electricity grid and cars have mostly existed in a single, unchanged format. But as we move beyond lithium-ion technology, a new generation of cheaper and more powerful batteries will completely rejig the power grid and usher in an age of electrically powered transportation.

Electric cars will replace imported oil with domestic electricity, use as little as one-fourth of the energy of petrol-driven cars per kilometre, emit significantly less carbon than conventional cars and, like mobile phones, provide a platform that supports apps that can do far more than just move the vehicle.

When both the drivetrain and the computer controlling it are electronic, it will be easier for them to communicate, so smart electric cars will know the traffic ahead, decide when to change lanes or choose alternative routes, find and pay for parking, program your driving calendar so that your itinerary saves energy and time, schedule the next charge based on your driving day and, ultimately, take over driving completely, doing it more effectively and safely than humans while freeing drivers for more productive and interesting tasks. Speciality cars for uses such as commuting, weekend errands, family events or long holidays may replace today's general-purpose cars. For many, car ownership itself may be replaced by renting or sharing on a need-to-use basis. Transportation will be personalized in the same way that mobile phones have personalized communication and information.

The energy-storage revolution will also shake-up the electricity grid. Access to adequate amounts of cheap energy storage will break the constraint that power must be generated at the same rate that it is used. Instead, we will have a 'bank' for electricity that can accept deposits and withdrawals at any time. Such flexibility is essential if renewable electricity is to become widely deployed.

Inexpensive energy storage will allow customers to 'draw off' electricity when it is cheap, such as in the middle of the night, and store it until they need it. And if the customer has local power generation, such as a solar panel on the roof, smart technology will switch between drawing power from the grid, the solar panel or local batteries and storing electricity, thus improving efficiency and lowering the cost of electricity use.

Currently, excess energy generated by local solar panels is sold back into the grid, and the homeowner uses grid electricity when the sun is not shining. But in the future, power from solar panels could be stored locally and used later — so the homeowner might rely much less on the main grid, instead relying on their own 'microgrid'. A neighbourhood of homes sharing a large common battery would constitute an even more

cost-efficient microgrid thanks to the economy of scale.

Commercial buildings, university campuses, factories and military bases would have different needs, requiring microgrids designed with their own special mix of storage, generation and use. Such diverse microgrids could serve their customers much more effectively than the one-size-fits-all approach of centralized utilities without storage.

A distributed, interacting network of microgrids would enable higher reliability, flexibility and resiliency, as well as security. A thunderstorm, hurricane or terrorist attack would not disable power for the entire network. Instead, surviving microgrids would share stored power with power-less neighbours.

How big will these changes be in energy and financial terms? In the United States, personal electronics account for about 2% of energy use, and make up most of the lithium-ion battery market, which is worth between US\$15 billion and \$20 billion; transportation and the grid

together account for nearly 70% of electricity use. If half of this transportation and grid energy was channelled through storage this would create a market of more than 15 times the current size.

What would it take to achieve these game-changing outcomes? One innovation: high-performance, inexpensive electricity storage. Simple estimates suggest that improvements in performance and cost of around fivefold are needed to enable an inexpensive electric car with a range of hundreds of kilometres, the replacement of fossil-fuel power plants with wind, solar and stored electricity, and the installation of compact, distributed storage in urban areas where land is expensive.

These fivefold improvements will not come from incremental advances in today's lithium-ion technology; they will require conceptual innovations and qualitatively different approaches that

go beyond lithium-ion technology. There are many promising avenues, including multivalent batteries that use doubly or triply charged working ions in place of singly charged lithium, the replacement of intercalation at the anode and cathode with covalent chemical reactions, as in lithium-sulfur batteries, and a host of flow-battery concepts based on high-voltage organic electrolytes and polymer active materials.

Although the societal pay-offs of electrified transportation and smart storage on the grid are substantial, the ultimate path to their development is uncertain and the risk of failure is high. Greater efforts by research organizations that can tolerate that risk, such as universities and national labs, are needed to identify and develop the most promising opportunities for next-generation energy storage. As the winning technologies emerge, the private sector will engage and deliver them to the public. As was true of the lithium-ion battery at its introduction in 1991, the challenges and opportunities are vast, rich and mostly unexplored. ■

**CHEAP ENERGY
STORAGE WILL
BREAK THE
CONSTRAINT
THAT POWER MUST
BE GENERATED AT THE
SAME RATE THAT
IT IS USED.**

George Crabtree is director of the Joint Center for Energy Storage Research at Argonne National Laboratory, Illinois, and at the University of Illinois at Chicago.
e-mail: crabtree@anl.gov



Smartphones are ubiquitous; they owe most of this success to the lithium-ion batteries that power them.

LITHIUM BATTERIES

To the limits of lithium

Researchers are developing a type of battery that has ten times the power of conventional batteries.

BY ERIC C. EVARTS

There is one major reason why you are able to carry around a powerful microcomputer in your pocket. Lithium-ion batteries have been credited for revolutionizing communications and transportation, enabling the rise of super-slim smartphones and electric cars with a practical range.

These innovations were possible because lithium-ion batteries can be much smaller and lighter than the previous generation of nickel-cadmium batteries, but still provide the same power. Better still, lithium-ion batteries retain their charge for longer and are composed of much less toxic materials.

As the lightest metal on the periodic table, and the one most eager to shed its electrons, lithium is the ideal element to make powerful, portable batteries. It can do the most work with the least mass and the fewest chemical complications.

But the development of lithium batteries was fraught with difficulties. The first versions — developed by the Texas-based oil company Exxon in response to the energy shortages

during the 1970s oil crisis — were not rechargeable and used lithium compounds that created toxic by-products in the electrolyte, unlike later lithium-ion batteries. They went on to power the first generation of digital watches, but early prototypes were ticking time bombs. Gases from the electrolyte could build up inside the battery and burst into flames as soon as they made contact with air.

Over the next 30 years or so, progress in lithium battery development had regular setbacks owing to battery fires and angry customers. Constant recalls gave lithium batteries a bad name, and sceptics assumed that they would never be safe enough for the mass market.

The promise of lithium batteries to transform the way society uses energy, however, is so powerful that it has created a gold rush among scientists, engineers, venture capitalists and entrepreneurs to tame their volatility.

“Modern society is completely dependent on fossil fuels, so there’s a huge incentive to find a replacement for the internal combustion engine,” says John Goodenough, a solid-state physicist at the University of Texas at Austin who is widely seen as the father of today’s

lithium-ion batteries. “We have to find a way to liberate society from that dependence. What we’re talking about is to be able to get electric energy from sun and wind, instead of from coal. But that’s not feasible unless you have storage.”

Goodenough is credited with three of the four major breakthroughs that led to the widespread success of lithium-ion batteries. In the late 1970s, he developed cathodes containing lithium cobalt oxide, which still power most personal electronics today (K. Mizushima *et al. Mater. Res. Bull.* **15**, 783–789; 1980). Later, he and a battery materials researcher, Michael Thackeray, followed this up with lithium manganese oxide cathodes, which power most electric cars today as well as some medical devices (M. Thackeray *et al. Mater. Res. Bull.* **18**, 461–472; 1983). In the 1990s, Goodenough carried on working after his retirement and developed an even cheaper and more stable cathode made of lithium iron phosphate, which is now widely used in power tools.

These chemistries are safer than the first generation of lithium batteries because the electrodes contain no free lithium. Instead, a chemical latticework tightly binds lithium ions



A researcher tests prototype electrodes as part of efforts to find ways of making batteries last longer and charge more quickly.

to complex metal oxide crystal structures in the positive cathode (where electrons and lithium ions travel as the battery is powering a device). The lithium ions move directly through the electrolyte and do not react with other elements.

THE LITHIUM LANDSCAPE

As far as lithium batteries have come, today's technology is still limited. Many mobile phones cannot make it through a day without being recharged. Most electric cars can travel for only 160 kilometres or less before they have to be plugged in for hours to recharge. The Tesla Model S offers a market-leading range of more than 300 kilometres, but at a cost: prices start at US\$71,000.

Battery scientists and engineers have been making batteries 5–10% more efficient every year for the past 25 years, says George Crabtree, a materials scientist at Argonne National Laboratory in Illinois. But he worries that this pace of progress may be slowing. “Unless a brand new oxide is discovered, that’s where we are,” he says.

“Where you are going to pull the next rabbit out of the hat, we’re not sure.” Even Goodenough plays down the impact that the technology has had. “We don’t really have electric cars yet,” he says. “You’ve got to lower the cost, make them safer and lengthen the driving range so you don’t give people range anxiety.”

The trick to developing batteries that last longer, charge more quickly and cost less is to pack as many lithium ions into each of the electrodes as possible, and get them to flow back and forth between the anode and cathode as quickly as possible, without letting them flow out of control.

For example, Goodenough’s former partner,

Thackeray, who now works on electrochemical storage technologies at Argonne, developed a crystalline molecular structure known as a spinel to transport lithium ions safely into the cathode as the battery discharges. The ions can only flow through channels in the spinel, making it highly stable. The more tightly that scientists can control lithium ions in structures such as these, the more stable the battery becomes. The technology does have a significant drawback, however: the less freely that lithium ions can move, the less power that the battery has. And as the transition from nickel–cadmium to lithium-ion batteries proved, boosting power is the best way to reduce size, weight and cost.

But it may not be the only way. Ping Liu, who directs research into advanced materials for energy efficiency and storage at the Advanced Research Projects Agency–Energy (ARPA-E), part of the US Department of Energy in Washington DC, says that improving safety and stability can also make batteries smaller and less expensive. That is because such advances would allow engineers to eliminate the electronic controllers, armour plating, cell insulators and cooling systems that today’s electric cars need to protect the batteries.

POSITIVE AND NEGATIVE DEVELOPMENTS

The most substantial progress so far has occurred at the cathode, the positive pole that absorbs lithium ions as the battery provides its power. Scientists have been working on making thinner layers of cathode material using nanomaterials, such as carbon, with the goal of speeding up the chemical flow of lithium through the battery by shortening the distance that ions have to travel. Some scientists are working on material as thin as a single atom. But efforts on this front have still not produced the advantages hoped for. “No one has managed

to do it well,” says Crabtree. “It’s difficult to make the layers come out even at such a small scale.”

The next breakthroughs are expected to come at the other end of the battery, from better anodes, Crabtree says. Anodes store the lithium ions when the battery is charged and send them to the cathode as the battery releases power. When the Japanese electronics giant Sony introduced carbon anodes to replace more troublesome lithium metal anodes in the early 1990s, the batteries lost some power. Now engineers want to get it back.

One big problem with today’s graphite anodes, and the lithium metal anodes that preceded them, is that the lithium ions returning to the anode when the battery is being charged do not coat the surface evenly. Instead, they grow like tree limbs in tiny crystalline structures called dendrites.

“Carbon can only accept lithium at a given rate,” says materials scientist Nitash Balsara at the University of California, Berkeley. “If you try to send lithium [through the battery] too fast [while charging], the lithium doesn’t really go into the graphite, it sticks on the outside. It becomes a safety hazard.” And the smaller the battery, the easier it is for dendrites to grow all the way across the electrolyte and contact the opposite pole, shorting out the battery, Goodenough says.

Permeable membranes called separators are used to prevent contact between the electrodes, and thus stop short circuits, while allowing the flow of electrolyte. But dendrites can break off and block pores in separators, shortening the life of the battery.

MORE POWER

Anodes can be made from silicon, which can hold up to ten times as much lithium per gram as graphite and therefore generate more power.

ARGONNE NATL LAB

But silicon poses its own problem: it expands to more than three times its normal size when the battery is charged and the anode is filled with lithium ions. This swelling breaks down the electrical bonds in the anode and stops the battery from working. It can also break the adjacent parts of the battery, such as the separator and even the battery case, and thus cause a fire.

Yi Cui, a materials scientist at Stanford University, California, who has been developing lithium-ion batteries for 15 years, is one of the scientists working on thinner electrode materials. He is developing silicon nanowires that stick up from the anode like fibres from a carpet and do not break the electrical bonds when they swell. But he says that the technology is still five years from commercialization. He is also experimenting with ways to improve graphite anodes, using two-dimensional graphene to absorb lithium more quickly while charging. But he says that this work too has a long way to go.

The ideal would be to return to a pure lithium-metal anode. "Forget silicon, if you can go straight to lithium-metal, that is the end goal," says Liu.

Compared with graphite anodes, lithium-metal anodes can absorb ten times more lithium ions while charging — and without silicon's swelling problem. Such a battery could reach a key performance metric for use in electric vehicles: delivering the 300 watt-hours of energy per kilogram that is needed to enable an electric vehicle to go the same distance on one charge that a petroleum-powered car can on a full tank. But this milestone would require other safety advances, such as a solid electrolyte or better separators, to allow charging without dendrite growth.

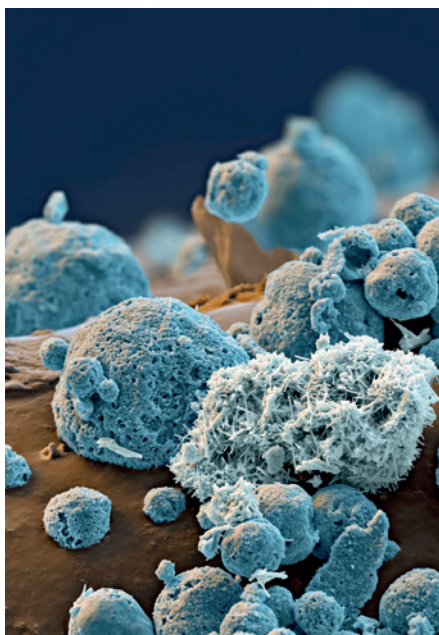
"I just don't see lithium metal with a liquid electrolyte as a commercial possibility," says Balsara. "That looks like TNT to me."

SOLID ELECTROLYTES

Solid electrolytes could bring the biggest breakthroughs yet (see page S8.) "If you could get rid of a liquid electrolyte, then you can get rid of all the combustible elements," says Liu. That would address the primary safety concern of lithium batteries. Liu's engineers are even designing a battery so strong and safe that it could ultimately form part of a car's body and absorb the impact in a crash, he says.

But solid electrolytes bring trade-offs, too. "They don't conduct electricity nearly as well as liquid electrolytes, so charging time goes up, and power goes down," Balsara says. And Goodenough adds: "I don't think you can get a solid-state battery with long life."

Goodenough aims to bridge the gap by developing a partial-solid electrolyte using a solid material next to the anode and a liquid next to the cathode. "It will achieve the same benefits as a solid electrolyte with longer life," he says, "because the anode material won't break down."



Lithium ion is used in rechargeable batteries.

He has built a test battery in his laboratory, but says he is still having problems with dendrite growth.

Liu says that the most promising developments are coming from scientists working on ceramic electrolytes. He points to an ARPA-E-funded project at the University of Maryland that has demonstrated a solid-state battery with a solid lithium anode that works using a ceramic electrolyte.

"Solid glass and ceramic electrolytes have significantly higher conductivity than plastic polymer," says Balsara, "and the conductivity is high enough that you might not have to sacrifice so much power."

Others working on new types of electrolytes for lithium-ion batteries include Yet-Ming Chiang, a materials scientist at the Massachusetts Institute of Technology in Cambridge, who helped to develop the lithium iron phosphate battery. He showed a new type of flow battery this summer that uses a semi-solid 'peanut butter-like' electrolyte. It solves the problem of how to get ions to flow quickly through a solid electrolyte by turning the battery design on its head, using liquid electrodes that are pumped over the semi-solid electrolyte. He says it could dramatically reduce the cost of producing lithium-ion batteries. Chiang has started a company called 24M, based in Cambridge, Massachusetts, to commercialize the invention, but it remains unproven.

NOT JUST IONS

As nanotechnology improves electrodes and electrolytes, the search for more power is bringing scientists full circle, back to batteries that use pure lithium rather than lithium ions. Lithium-sulfur and lithium-air batteries are not categorized as lithium-ion batteries because the lithium reacts in the

electrolyte to form other compounds rather than simply flowing through the electrolyte and not reacting with it.

Lithium-sulfur batteries, similar to those batteries that Exxon experimented with in the 1970s, can store up to ten times the energy of a lithium-ion battery by weight. The problem is that the electrochemical reactions involved consume the sulfur and create other substances that dilute the electrolyte. Both processes prematurely kill the battery. "Ideally, you'd like an electrolyte that only transmits lithium," says Balsara. "Anything else it does is a problem."

To solve this complication, Cui, Crabtree and others are developing nanomaterials to encapsulate the sulfur. Cui says that his lab has demonstrated a lithium-sulfur battery that can last through 500 to 1,000 charge-discharge cycles. That could be enough for a mobile phone or laptop, he says. But its energy capacity is still too low, notes Cui, and a commercial prototype may be five years away.

Lithium-air batteries could turn out to be the ultimate lithium battery in terms of power, weight and cost. With a lithium-metal anode and a gaseous oxygen cathode, a lithium-air battery could store as much energy as a lithium-sulfur battery at even less cost, and potentially with less weight.

If the materials sound simple, however, the battery is not. For one thing, 'air' is a misnomer. The cathode has to be pure oxygen, without any of the moisture or carbon dioxide found in air. The systems to purify, pump and store the air add 30–70% to the battery's weight and size, says Crabtree. Although on paper lithium-air can deliver ten times the energy density of lithium-ion, "it will never achieve the factor of ten that you get from the back of the envelope," he says.

On top of that, the cathode oxidizes any organic electrolyte over time, diminishing the life of the battery.

Worst of all, because the reaction turns lithium into non-conductive dilithium oxide, the battery is difficult to recharge. Cui says that researchers have demonstrated a reasonable number of charges and discharges of lithium-air test batteries by encapsulating dilithium oxide on a conductive substrate, but he estimates that lithium-air batteries are at least ten years away from commercialization.

Their development may sound difficult, but better lithium batteries could deliver huge pay-offs. Beyond electric cars that could go 300 kilometres, they could provide the breakthrough that makes renewable energy such as solar and wind power ubiquitous, and finally break the developed world's dependence on fossil fuels. ■

Eric C. Evarts is a freelance writer in Ridgefield, Connecticut.



Some electric cars are already using solid-state batteries, but a battery for the mass market is years away.

TECHNOLOGY

A solid future

Swapping the liquid electrolyte in batteries for a safer solid-state interior is bringing electric cars to the mass market.

BY JIM MOTAVALLI

During the first eight months of 2015, a row of five electric vehicles sat parked in Indianapolis, Indiana, available for test drives. The compact vehicles are part of the first US wave of the ambitious BlueIndy electric car-sharing programme from the Bolloré Group, headquartered in Paris, which is already running similar large-scale schemes in London and the French capital.

These Bluecars are a technological milestone: with unique 30-kilowatt-hour battery packs designed by Bolloré, they are the first plug-in vehicles on the road to have solid-state batteries, rather than the liquid-electrolyte batteries that many other electric vehicles use.

BlueIndy president Hervé Muller says that the plug-in vehicles' batteries have been performing well. And the safety record for Bolloré's batteries is good. "We haven't had an issue in Indianapolis, nor in Paris with the 3,000 cars in service there that have driven more than 10 million miles," Muller says. Bolloré's BlueIndy batteries, although groundbreaking, are an early iteration of solid-state technology and not yet ready for

the mass market. Scientists, however, are working on more advanced cells.

PROMISES AND CHALLENGES

The lithium-ion batteries in most electric cars consist of a negative — usually graphite — anode, a positive cathode and a liquid electrolyte. When these batteries release power, lithium ions move from the anode, through the electrolyte, to the cathode. The more conductive the electrolyte, the better the battery performs.

The promise of lithium-ion solid-state batteries is that they will replace the heavy and sometimes dangerous liquid electrolyte — which in car batteries can be volatile at high temperatures when the battery is charged or discharged quickly or when packs are damaged in accidents — with a lighter, more versatile solid alternative.

Although finding a solid electrolyte with conductivity that is comparable with liquids has been a challenge, the advantages are many. The batteries are safer because flammable components have been removed. They deliver more power because solid electrolytes mean that the carbon-based anodes can be replaced with lithium metal, which has a higher energy density and cycle life,

with less weight and cost. And without the need to package the liquid electrolyte safely, solid-state batteries can be made in more versatile shapes (even thin films), reducing manufacturing costs. This could make electric cars a more enticing proposition, with longer ranges and a lower purchase price.

"Imagine batteries that do not catch fire and do not lose storage capacity. That is the promise of solid-state batteries," says Gerbrand Ceder, a materials scientist at the Massachusetts Institute of Technology (MIT) in Cambridge.

Solid-state batteries are not quite ready for mass-market vehicles yet — the Bluecars' batteries need to be warmed up and can power a very small car for 240 kilometres. Scientists are struggling to find solid materials with conductivity similar to that of liquids, as well as working to increase cycle life and longevity, and to improve the solid electrolyte's ability to operate at ambient temperatures.

Yan Eric Wang, a materials scientist who worked with Ceder at MIT, and his team may have found the ideal solid electrolyte (Y. Wang *et al. Nature Mater.* **14**, 1026–1031; 2015). "We found a framework that could lead to identifying electrolyte materials with high ionic conductivity," says Wang. The research, in partnership with electronics firm Samsung, pointed to a class of compounds of lithium, phosphorus, germanium and sulfur called super-ionic conductors.

A solid electrolyte allows batteries to switch to lithium metal anodes, says Jeff Chamberlain, a battery researcher at the Joint Center for Energy Storage Research at Argonne National Laboratory in Illinois. "With a solid-state battery you open up the field of metallic anodes, and make possible a big jump in energy density. That would be a game changer."

To be commercially viable, solid-state batteries need to work reliably for years in a tough automotive environment. And durability is an issue for researchers. Lithium metal, although excellent at storing large amounts of energy at low volume, is very reactive. And it is prone to forming 'dendrites' — tiny lithium spurs that degrade battery performance and can cause a short circuit if they reach the cathode. With conventional lithium-ion batteries, a chemical separation layer guards against dendrites.

Sam Jaffe, managing director of Cairn Energy Research Advisors in Boulder, Colorado, says that solid-state researchers are working with additives in the electrolyte carrier, as well as ceramic shields, in an attempt to block dendrite formation.

PHONES FIRST

In the short term, as solid-state science is evolving, the durability issue may not matter as much in applications such as mobile phones, because consumers tend to switch to newer technology in a year or two — whereas cars are kept for a decade or more. "People turn over portable electronics very quickly," Chamberlain says, "so the

➔ **NATURE.COM**

To read more about better batteries visit:
go.nature.com/xm5vs6

ability to inject new technology is higher.”

SolidEnergy Systems, headquartered in Waltham, Massachusetts, uses technology developed at MIT and is targeting smartphones. The firm developed an improved electrode material for solid-state batteries that replaces the usual graphite with a very thin film of lithium-metal foil. The electrode's ability to store lithium seems promising (go.nature.com/b3xkhh), and the higher the storage capacity, the more energy the battery can deliver.

Something of a hybrid, SolidEnergy's cell maximizes efficiency by using both a solid and an ionic liquid electrolyte — and works at room temperature. A high-performance car battery is promised within the next four years.

INTERNATIONAL INVESTORS

Commercial solid-state cells for mass-market electric cars are at least a decade away, but they are coming, researchers say. Solving the significant technical hurdles has become a central focus for a group of start-up companies and research labs. Most companies, even those that have generated interest from vehicle manufacturers, are still in the developmental stage. Cosmin Laslau at Lux Research, headquartered in Boston, Massachusetts, says that today's best lithium-ion batteries are approaching 250 watt-hours per kilogram (the early Nissan Leaf had only 140 watt-hours), an impressive measure of energy density, but they will struggle to reach 350 — a performance goal set by the US Battery Consortium. If solid-state batteries can reach 350 “that's a pretty good sustainable advantage over traditional lithium-ion,” he says.

This could be good news for the electric vehicle industry. Donald Sadoway, a materials chemist at MIT, says that achieving such high energy densities is key to widespread adoption of electric vehicles. “If we had batteries with 350 watt-hours per kilogram we'd have EVs with 350 miles of range, and that's the end of petroleum,” he says.

But scaling up from a cell that generates results in a bench test to a durable, roadworthy battery pack will be a long road. Menahem Anderman, a physical chemist and chief executive of Total Battery Consulting in Oregon House, California, authored the California Air Resources Board's 2000 report on battery technology. “I have not seen any indications of a significant breakthrough or signs that they have a technology likely to find its way into the mass-produced vehicle inside the next ten years,” he says. “A new development often shows improvement in one or two areas but experiences difficulties in others.”

Chamberlain agrees, and he stretches the timeline a bit further. “To be in showroom cars ten years from now, the solid-state cells would essentially have to be leaving the laboratory now, and I don't think that's happening,” he says.

Still, vehicle manufacturers are interested, whatever the timeframe, as evidenced by a rash of investment and acquisitions over the past few

years. In 2014, Volkswagen bought a 5% stake in QuantumScape, an electronics firm in San Jose, California, that is hoping to benefit from developing solid-state technology that could triple the range of its forthcoming electric cars.

Car manufacturer Toyota says that cells that it has developed with double the energy density of today's alternatives could provide electric vehicles with a range of 480 kilometres on one charge. It has built prototype cells and even a small scooter powered by the batteries. Positive results have been reported for solid super-ionic electrolytes (N. Kamaya *et al.* *Nature Mater.* **10**, 682–686; 2011). But the company says that cell-energy density is still far below the potential, and production of the batteries is not expected to start until the early 2020s.

Battery-development firm Seeo, based in Hayward, California, which was acquired in August by German auto supplier Bosch, is also making solid-state cells.

Ulrik Grape, a vice-president at Seeo, says that the company's cells represent “a very durable technology with good cycle life. They're as good or better than current lithium-ion.” He adds that Seeo packs could be half the weight of those in the current Nissan Leaf — the leading battery electric car worldwide in terms of sales. The company also says its cells achieve 350 watt-hours per kilogram in the lab, but the real-world performance could be less.

One more player, Solid Power, which is based in the Colorado Technology Center in Louisville and is a spin-off from the University of Colorado at Boulder, says it has made lab-scale cells that have reached 400 to 500 watt-hours per kilogram, and up to 500 charging–discharging cycles of durability. Company founder Doug Campbell says that Solid Power is targeting the automotive sector, although its first market may be the armed forces, for use in communications equipment. “Troops in the field can carry 60 pounds of batteries, and if we can cut that in half, it's a strong value proposition for the US military,” he says.

For competitive reasons, many of the lab-stage battery companies are keen to emphasize the positive aspects of their research, but not all are forthcoming with sample cells or technical details. Sakti3, based in Ann Arbor, Michigan, and headed by former University of Michigan engineer Ann Marie Sastry, is promising greatly improved energy density and an improved manufacturing process, with weight and cost savings, from its thin-film cells that operate at ambient temperatures. Sastry says that the company, with investments from General Motors and UK-based Dyson, has demonstrated that it can double the density of conventional lithium-ion at the lab scale.

But Sakti3 has released little information



Lithium-ion batteries have led to fires in some aircraft.

about its chemistry or its results, so its claims are difficult to verify. Laslau estimates that Sakti3 has reached a fairly impressive 300 watt-hours per kilogram, but Sastry says that the company is not disclosing energy-density results just yet.

Cost to manufacture is another important factor. “We do think that solid-state technology will enable much better performance, at lower cost, than the incumbents,” Sastry says. Her company is aiming for cells that cost just \$100 per kilowatt-hour, which means around \$2,400 for an average-sized electric-vehicle pack, plus costs for packaging, transportation and other factors. Nissan currently sells its lithium-ion packs for \$5,499 in the United States, although it may be losing money on every sale. Asked if her cells could enable the goal of a 450-km, \$25,000 electric car, Sastry says: “At least.”

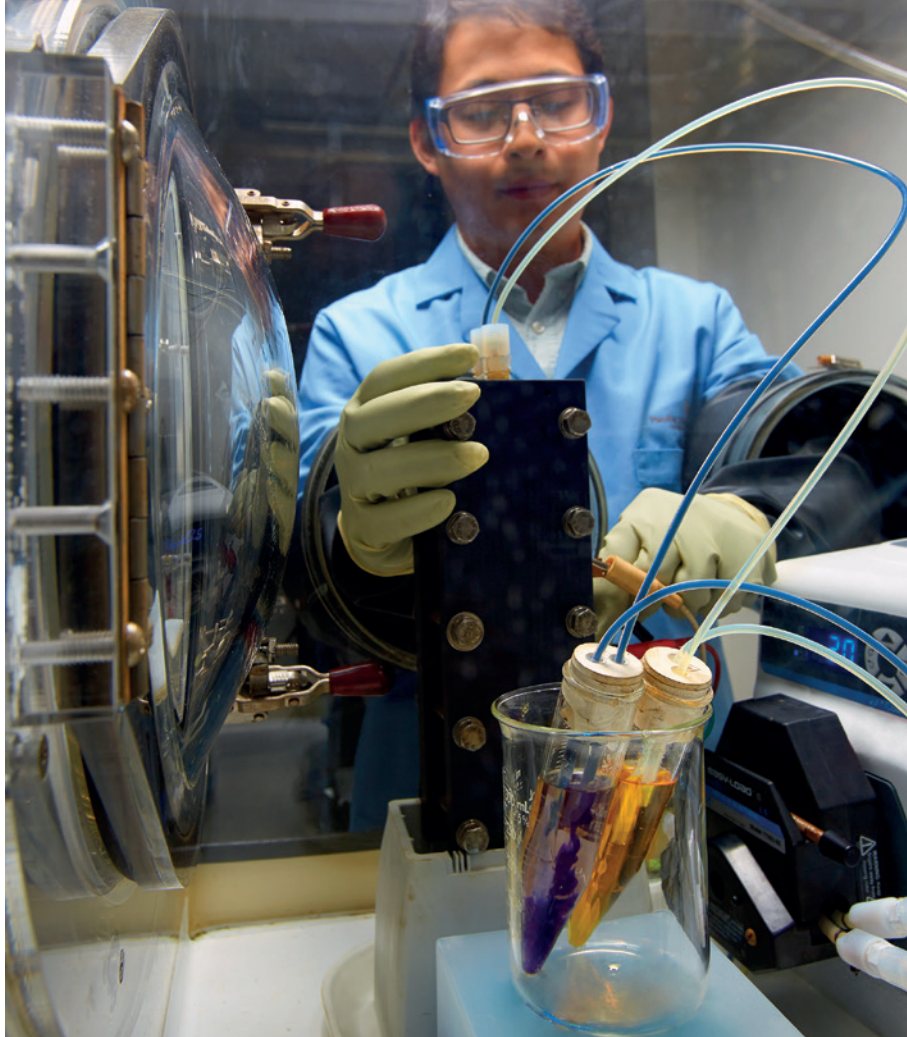
Scaling up to that level, Sastry admits, will be challenging because of the need to invent manufacturing processes. But “everything we have done to date can in principle be done at scale.”

The 30-kilowatt-hour solid-state lithium metal polymer battery packs in the Bluecars are here now, but the 240-km range is not a huge improvement on current technology. BlueIndy's packs require heating above ambient temperature, and that uses some of the battery's energy and reduces range.

Solid-state-battery researchers will surely hit some as-yet unforeseen roadblocks on the way to a commercially durable cell. For cars, conventional lithium-ion probably has at least a decade of dominance remaining, although Sadoway believes that “the rush to deployment would be very, very fast” if stable, high-density solid-state cells were developed.

Solid-state batteries, Sadoway adds, “would be preferable, because they're a lot safer”. The lithium-ion fires that have occurred on some aircraft show what can happen when you rush to scale up battery technology without doing your homework first. ■

Jim Motavalli is freelance writer based in Fairfield, Connecticut.



A researcher at the Pacific Northwest National Laboratory in Washington working on flow batteries.

ELECTROCHEMISTRY

Liquid assets

Flow batteries, which release electricity through fluid-based reactions, could revolutionize renewable-energy storage.

BY NEIL SAVAGE

When state officials flipped a switch earlier this year at an engineering company in Pullman, Washington, they shone a light on one possible future for energy storage. That switch activated the latest type of flow battery, the largest in the Western Hemisphere. Rechargeable flow batteries, which store energy in tanks filled with liquids, have the potential to be cheaper than their conventional, solid cousins. They are also more adaptable to the needs of electrical grids, which are starting to rely on intermittent sources of energy such as wind and solar cells.

The Pullman flow installation, made up of big white boxes that together take up roughly the same amount of space as two tractor-trailers, stores 4 megawatt-hours of energy. That is more than enough to run four average homes for a month. Although the installation holds only a

fraction of the power the grid will require, it is introducing a new generation of energy-storage technology.

If wind and solar power are ever to provide a significant portion of the world's electricity, new ways will be needed to store that energy. Existing batteries such as lithium ion and lead acid do not provide the necessary combination of long-term energy storage and rapid delivery of energy — just think how quickly a lead-acid car battery can be drained by a driver trying repeatedly to start a car on a cold day, or the overnight charge that an electric vehicle needs.

Flow batteries could provide an alternative. They can store energy for a long time, but provide it quickly when needed; they are liquid-based, so inherently safer than conventional batteries; and because the energy-storing liquids are kept in external tanks, changing their storage capacity is relatively

simple. Most importantly, if researchers can develop the right combination of chemistries, flow batteries could be much less expensive over their lifetime than existing batteries.

GO WITH THE FLOW

A flow battery is a type of fuel cell that consists of two tanks, each containing an electrolyte made of some sort of energy-storing material — a metal or a polymer — dissolved in a liquid. One liquid is the negative side of the battery, the other the positive side. The liquids are pumped through a stack that contains positive and negative electrodes and a membrane that keeps the liquids from mixing, but allows ions to cross. The liquids undergo a reduction and oxidation, or redox, reaction, transferring electrons from the negative to the positive tank (see 'Renewable energy storage').

The most common version uses the transition metal vanadium (V) dissolved in acid. During discharge, ions of V^{2+} on the negative side of the battery are oxidized to V^{3+} . The electrons given up by the vanadium ions flow out to an external circuit, providing a current to whatever the battery is powering. The electrons travel to the positive side, where they latch onto a different ion, V^{5+} , reducing it to V^{4+} . During charging, this redox reaction is reversed, removing electrons from the positive side to recharge the negative side.

This design offers some inherent advantages. For one, unlike in a conventional battery, the power capacity and the energy capacity are separate. Energy is determined by the volume of the electrolyte, whereas power is controlled by the area of the electrode stack. "If you need to increase your capacity, you can easily just buy a new tank and fill it up with the chemicals," says Adam Weber, a chemical engineer at the Lawrence Berkeley National Laboratory in California. For higher powers, you can simply run the electrolyte through more or larger stacks.

Storing the electrolytes separately is also safer. In conventional batteries, Weber says, "you're trying to store a lot of power and energy in a little box". This design leads to a risk of catastrophic discharge of stored energy that is absent in flow batteries. Because many flow batteries are water-based, they are not likely to catch fire, unlike the lithium-ion batteries that have destroyed mobile phones and grounded airplanes.

Another problem with conventional batteries is that the diodes swell and shrink as ions pass back and forth through them, eventually leading the materials to fracture and fail. Liquids do not crack, so the electrolytes in flow batteries can last indefinitely.

CLEANER TRANSPORT

Although most researchers are developing flow batteries with an eye towards grid storage, or to accompany home-based solar energy, there are also efforts to build flow batteries that work in electric vehicles. Because a flow battery can be

recharged simply by exchanging the liquid in it, owners of electric vehicles will not have to worry about stopping to charge their cars overnight. “Range doesn’t become so important, as long as there’s a filling station somewhere,” says Carlo Segre, a physicist at the Illinois Institute of Technology in Chicago. He has a grant from the Advanced Research Projects Agency–Energy (ARPA–E) to develop a flow battery for cars that he says could have a range of 800 kilometres or more (Tesla Motors says that the batteries used in its Roadster vehicle have a range of about 640 km). Segre’s battery uses nanoparticles of nickel and nickel hydroxide suspended in a potassium-based electrolyte.

The challenge in making flow batteries viable lies in finding designs and chemistries that provide good long-term storage, desirable current and voltage characteristics, a long lifespan and a competitive price. The Pullman battery, and a few higher-power batteries — two installed in China, three in Japan — are all based on vanadium chemistry, which provides both high energy and power.

Usually, the vanadium ions are dissolved in sulfuric acid. But there is a limit to how many vanadium ions the acid can hold, capping how much energy a given volume of electrolyte can contain. On top of that, the chemistry only works between 10 °C and 40 °C. Outside that range, the vanadium precipitates out of the acid, as either a powder or a gel. “Once it precipitated out it would jam your pumping system. Your battery would be dead,” says Wei Wang, a materials scientist at the Pacific Northwest National Laboratory in Richland, Washington. Wang and his team developed a solution of sulfuric and hydrochloric acids that can handle temperatures from –5 °C to 60 °C, and holds about 70% more vanadium than current systems¹. The Pullman battery uses this new electrolyte.

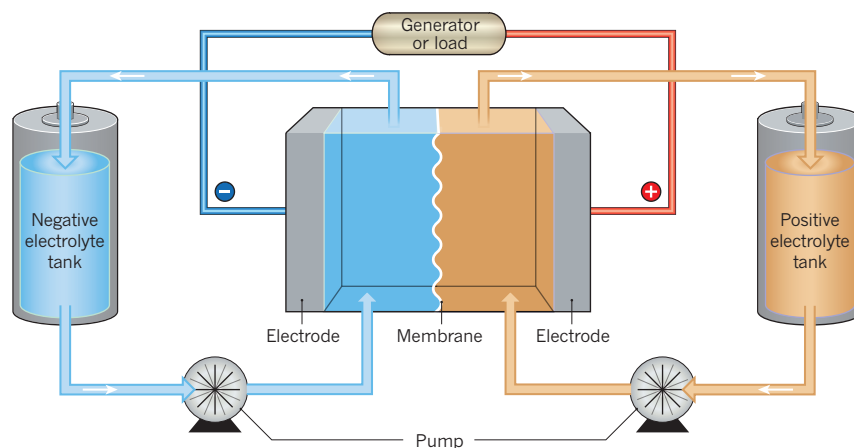
Even with this expanded range of operating temperatures, vanadium is not the ideal flow-battery material. It is rare and expensive — a 2011 estimate² from the energy-company-funded Electric Power Research Institute puts the cost of a vanadium redox flow battery at US\$3,000–3,310 per kilowatt. At that price, it would take about \$4-billion worth of vanadium batteries to provide California with the 1.3 gigawatts of storage it wants by 2020.

To find a cheaper alternative, Wang is working on zinc bromine batteries. Zinc is much cheaper than vanadium, and can carry more electrical charge. But Wang uses it as a solid, so one half of the hybrid flow battery does not flow and the volume of the zinc electrolyte cannot be changed. Still, he says, the hybrid flow battery allows more flexibility in system design than conventional devices, and its cost and performance may turn out to be appealing.

Instead of replacing a liquid with a solid, Weber made one side of his flow battery a gas. During discharge in his hydrogen bromine battery, gaseous hydrogen is oxidized at

RENEWABLE ENERGY STORAGE

Tanks of metals or polymers in liquid make up the positive and negative sides of a flow battery. The liquids are pumped through electrodes, where they undergo reactions to generate electricity when a load is connected.



the negative electrode, producing an ion that passes through the membrane and reacts with the bromine. The battery has reached some of the highest powers of any redox flow chemistry in the lab³. And because hydrogen is inexpensive, the system could be a winner economically.

COST REDUCTION

Lithium, seen as the main conventional competition for flow batteries, may have its place in these upstarts as well. Yet-Ming Chiang, a materials scientist at the Massachusetts Institute of Technology in Cambridge, is developing a lithium sulfur flow battery⁴. Researchers have been working on conventional lithium sulfur batteries, but the elements tend to react, creating polysulfides that migrate to the electrode and block the flow of current. What is bad for conventional batteries might be good for flow batteries. Chiang’s set-up uses sulfur as a positive electrode. During discharge, lithium ions from the negative electrolyte move to and react with the sulfur cathode, yielding polysulfides that are suspended in the electrolyte liquid. Charging the battery causes the polysulfides to dissolve, and the sulfur precipitates out, recreating the positive electrode. And because sulfur is readily available, the battery should be inexpensive. “The attraction of sulfur is that it really is essentially free as a material,” Chiang says.

His other innovation is to replace the flow battery’s current collector, which is typically a stationary mesh of carbon fibres that the electrolytes must pass through. The mesh can clog, so instead, Chiang created a gel of nanoscale carbon particles, which moves along with the electrolyte and acts as a sort of ‘liquid wire’ to collect the current. That provides a greater surface area to absorb the charges from the chemical reaction, while shortening the distance the molecules have to travel to give up their charge.

Another potentially affordable option could be organic materials. “We’re not going to run out of carbon, hydrogen and oxygen,” says Michael Aziz, a materials scientist at Harvard

University in Cambridge, Massachusetts. Aziz was inspired by plant photosynthesis to test quinones, a family of organic molecules common in plants. Two years ago, he demonstrated⁵ a metal-free flow battery using bromine on one side and a quinone molecule called AQDS — which occurs naturally in rhubarb — on the other. He has gone on to synthesize other versions of quinone, looking for the one that is least expensive and most workable. “Getting the overall cost down is a big deal, and doing it in an environmentally friendly way is even bigger. We think we have a fighting chance of doing both,” says Aziz. But the organic route is not without its pitfalls. Organic molecules tend to undergo many different chemical reactions, and the wrong ones could produce insoluble particles that gum up the battery.

In the near term, the batteries most widely used to store energy on the grid will be conventional lithium-ion batteries, Chiang says. But several types of flow battery offer combinations of efficiency, safety and cost that could allow them to displace conventional batteries. Meanwhile, if Segre and others develop workable flow batteries for electric vehicles, the fact that they can be refilled like a car’s petrol tank would make them an attractive alternative to today’s battery technology. Although researchers are still working to get just the right designs and chemistries, Aziz believes that flow batteries could be ready for a great leap forward. “In two years we’ll have something that’s at roughly the industrial use level,” he predicts. “I think it’s the next big thing.” ■

Neil Savage is a freelance science and technology writer in Lowell, Massachusetts.

1. Li, L. *Adv. Energy Mater.* **1**, 394–400 (2011).
2. Dunn, B., Kamath, H. & Tarascon, J.-M. *Science* **334**, 928–935 (2011).
3. Cho, K. T. et al. *J. Electrochem. Soc.* **159**, A1806–A1815 (2012).
4. Fan, F. Y. et al. *Nano Lett.* **14**, 2210–2218 (2014).
5. Huskinson, B. et al. *Nature* **505**, 195–198 (2014).



Processing of lithium-ion batteries for recycling at Retriev Technologies in Anaheim, California.

RECYCLING

Lazarus batteries

Battery recycling can be hard, energy intensive and uneconomic. But soon, dead power cells could be more easily resurrected.

BY ERICA GIES

Solar panels and wind turbines are becoming an increasingly common sight in many countries, and city dwellers are adjusting to the eerie hush of electric cars — nowhere more so than in Norway, where one-quarter of new vehicles sold in the first three months of 2015 were electric or hybrids. But these signs of progress towards a more-sustainable future rely on big batteries that are packed with chemicals and mined metals, and take large amounts of energy to manufacture.

Recycling can soften the environmental impact of batteries by reducing the energy required for their production, as well as the environmental harm caused by the disposal of hazardous battery materials and the mining of new ones. Sometimes, it can also provide an economic incentive, because valuable recovered materials can be sold on.

Unfortunately, most types of batteries are not being recycled, primarily because battery recycling is not required by law in most countries. Some US states have policies for certain battery types that encourage collection

and recycling, but they do not make battery recycling compulsory. A European Union law mandates recycling, but it is being phased in over time. That leaves economics as a motivation, but in general, batteries do not contain enough precious metals to make recycling economically viable. In addition, the recycling processes available today reduce extracted metals back to their original form, meaning that multiple processes must be repeated to build a new battery from them.

But a promising technology could provide a much simpler and less destructive way to recycle batteries in the future. And better planning at the design stage could help.

LEAD TAKES THE LEAD

At present, one form of battery is already widely recycled. Lead-acid batteries (which are ubiquitous in petrol-fuelled cars) have about a 98% recycling rate in most developed countries, primarily because lead is toxic and its disposal is heavily regulated. Furthermore, most car batteries are replaced by professional mechanics, who are plugged into the battery-recycling network.

Recycling lead-acid batteries also makes economic sense: it costs less than mining new lead, and because this heavy metal makes up about 65% of the battery's mass, it is easy to recover. The process saves money for the manufacturers and is lucrative for scrap dealers, says Joe Acker, who has worked in lead-acid battery recycling for 25 years and is now president of Retriev Technologies, formerly Toxco, a battery-recycling company in Anaheim, California.

LITHIUM-ION LAGS BEHIND

But whereas lead-acid batteries are almost always recycled, it is much less common for the lithium-ion batteries in mobile phones, portable electronics and electric cars to be recovered. In the EU, just 5% of the lithium-ion batteries sold in 2010 were recycled.

One reason for this low rate is that recycling lithium-ion batteries can be complicated. Lead-acid batteries have few materials and designs, says Linda Gaines, an analyst for the Center for Transportation Research at Argonne National Laboratory in Illinois, and breaking them apart to recover constituent parts is easy because the product is standardized across the market. Conversely, lithium-ion batteries have various chemistries and shapes. The cathode could include cobalt, nickel, manganese, phosphorus and iron along with lithium, whereas the anode is typically carbon, and the organic electrolyte contains a lithium salt. Copper, aluminium and steel are also used as the substrate for the electrode materials and for the housing.

Much of the lithium-ion battery recycling so far has been motivated by the extraction of cobalt from the cathode, along with, to a lesser extent, nickel and copper. The value of these elements makes recycling economically attractive. In particular, cobalt is expensive and difficult to obtain; one of its primary sources is the conflict zone in the Democratic Republic of the Congo.

However, cobalt is gradually being phased out as researchers find cheaper cathode materials. So newer batteries, including “most of those now being built for vehicles, are less attractive for recycling by current methods”, says Gaines.

Methods of recycling in commercial use today are based on metal refining. First and most widespread is pyrometallurgy, in which batteries are fed into a smelter and melted down to recover metals. In older lithium-ion batteries, cobalt might make up 18% of the battery. But the rest of the materials, including lithium and aluminium, end up in slag and are then buried or embedded in concrete.

The Brussels-based battery-recycling company Umicore uses pyrometallurgy to recover cobalt, nickel, copper and other metals from lithium-ion batteries and nickel-metal hydride batteries. However, profits from selling recovered metals are not a major driver of its

LEFT: UNICORE; RIGHT: D. BOSWELL/J. ALBERTS

recycling operation, says Maarten Quix, the firm's head of battery recycling. Rather, battery manufacturers pay the company to recycle their products. "Producers and collectors pay for the service in order to ensure that the credentials of 'green' products are solid, or to meet legislation targets," he says.

A second method, hydrometallurgy, is used by recyclers such as Retriev Technologies. In this process, cells are chopped up in water and separated into various product streams on the basis of what sinks, floats or hangs in solution. Chemicals are then added to the water, where they react with cathode components to make new compounds that are easier to separate.

DIRECT RECOVERY

But both pyrometallurgy and hydrometallurgy have limitations. Aside from their high energy and emissions footprint, many materials, including lithium and aluminium, end up in the waste stream after smelting, which means that more materials must be mined to build new batteries. "The model as it is now drains material away from batteries," says Steve Sloop, president of OnTo Technology, a research and development company in Bend, Oregon, that is working on recycling batteries.

Sloop developed a method called direct recovery. Rather than recovering basic elements, as with pyrometallurgy, or partially breaking down the molecular structure of compounds, as with hydrometallurgy, this method involves bathing the cathode in a soft chemical solution to rejuvenate it. The reactions are topotactic, meaning that structural elements of the cathode are retained, but its chemical composition is changed. In this way, cathode materials such as lithium that are not valuable enough to be recovered by pyrometallurgy or hydrometallurgy can be reused in batteries. It is a low-temperature, low-energy, low-emissions process with very limited waste, he says.

"It's really cheap to do this," says Sloop. Most of the cost savings come from not having to rebuild the cathode again, and direct recovery uses less energy than pyro- and hydrometallurgical processes. It also produces fewer emissions and waste, saving money on compliance with air-pollution regulations and disposal fees.

The process is a closed-loop system, says Sloop: "Aluminium, plastics, carbon, metal parts of the cell can all be reused." He calculates that because of this thriftiness, direct recovery requires less than 2% of the energy required for new metal refining. Nor does it depend on large volumes; recyclers can add modular plants as the volume of material ramps up.

Gaines says that she sees great promise in direct recovery, which is also being developed by Retriev. "I don't see how any of the other

"Aluminium, plastics, carbon and metal parts of the cell can all be reused."



Pyrometallurgical processes convert used batteries into metal alloy (left); Steve Sloop has developed a method that produces fewer emissions and waste (right).



processes that have been developed so far can produce a valuable product from the cathodes that don't have valuable elements in them," says Gaines. With direct recovery, she says, "the value lies in retaining the materials in their battery-ready state".

Although direct recovery is not yet being used commercially, early tests are promising, says Sloop. "The performance levels are as high as with new materials." OnTo Technology is testing a car-grade battery built with rejuvenated material from a worn-out electric-vehicle battery. "It's still cycling, 2,000 cycles later," he says.

POLICY PUSH

Until direct recovery has proved itself, top-down regulations can help to kick-start recycling programmes and drive innovation — as happened with lead-acid batteries. In the EU, a manufacturer is responsible for its product through its entire life cycle — especially collection, recycling and final disposal. The 2006 EU Battery Directive set a timeline for recycling progressively greater percentages of batteries and made the battery producers responsible for those efforts.

The programme is still in its early days — the directive's first target was 25% battery recycling by September 2012 and the next target is 45% by September 2016. The data so far are murky, says Gaines, so it is not yet clear how successful the directive has been in increasing recycling.

In the United States, there is no federal regulation for battery recycling, leaving a patchwork of state-level rules. Industry has made some effort to fill that gap. Since 1996, a voluntary programme called Call2Recycle, which is funded by electronics manufacturers, has recycled more than 38.5-million kilograms of small consumer batteries and mobile phones in the United States and Canada, but that is still just a tiny fraction of the total. Carl Smith, Call2Recycle's president, says that the organization recycles perhaps 2% of alkaline batteries

and possibly 12% of lithium-ion ones. British Columbia, Quebec and Manitoba in Canada have mandatory recycling programme, and in these places, Call2Recycle collects about 25% of what is sold, says Smith.

PLANNING AHEAD

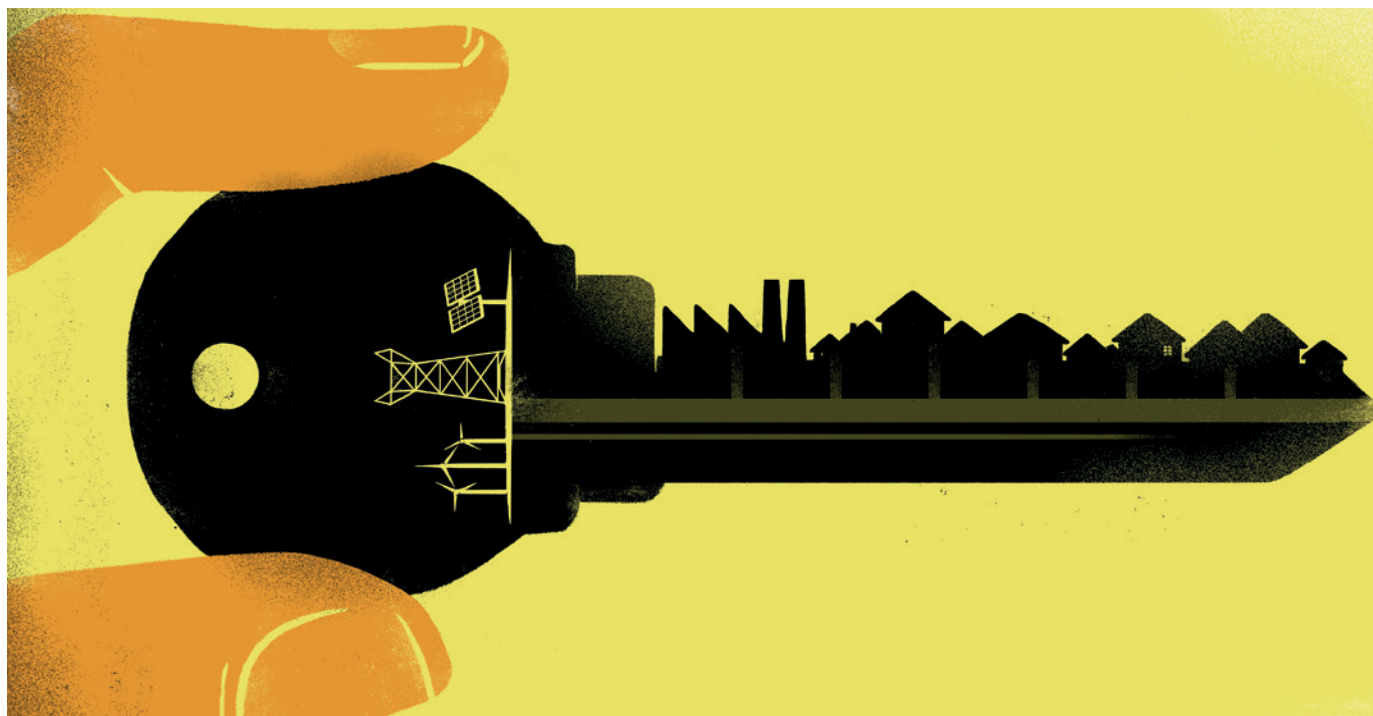
Prescient product design could help to increase recycling. Gaines has some ideas for strategies to facilitate recovery and reuse, such as using the minimum number of components, standardizing formats and materials, and avoiding toxic materials (such as cadmium, arsenic, mercury or halogens). She also recommends designs that more easily allow separation, such as nuts and bolts instead of welds.

Today's batteries are difficult and labour-intensive to disassemble because not much thought goes into their end of life, says Acker. More than 130 energy-storage companies operate in California, but end-of-life concerns are widely ignored by innovators and the venture capitalists who fund them. Instead, they prioritize performance and cost, the better to grab market share. San José State University in California offers a battery-focused master's degree in electrical engineering that has courses on various kinds of electrochemistry, manufacturing and technology, but "end-of-life considerations do not enter into it", says Shahab Ardan, the electrical engineer who oversees the master's programme.

That lack of foresight is unfortunate, says Gaines. "If you don't think about recycling early, you could develop some battery chemistry that's going to be completely intractable."

Nevertheless, she is hopeful about the future of battery recycling, particularly because direct recovery holds the promise of making it much more economic. Even if governments fail to wield policy sticks, economic carrots could finally deliver a greener energy future. ■

Erica Gies is a freelance science writer in Victoria, British Columbia.



ENERGY STORAGE

Power revolution

Electrical grids increasingly depend on intermittent renewable sources. To smooth the supply out, utilities companies are testing alternatives to storing energy in conventional batteries.

BY PETER FAIRLEY

It is 2025 and another sweltering summer's day in California. Millions of solar panels are soaking up the Sun's rays to power the air-conditioning systems that keep homes and offices throughout the state cool. The devices are working efficiently thanks to an intelligent conversation taking place between the appliances and the electrical grid. As clouds drift across the Sun, casting shadows, the air conditioners deftly increase or decrease their output in sync with the varying flow of solar energy. In areas where the demand for electricity looks as though it will overload the power-transmission lines, home air-conditioning units take it in turns to go offline for an hour. In other areas, where solar power threatens to exceed demand, hot-water heaters are turned on to absorb the extra energy.

This imagined future power grid demonstrates the same degree of flexibility that energy-storage advocates predict will occur with the widespread implementation of batteries, but there is no electrochemistry involved — software manipulates energy-consuming equipment so that most

electricity is used when it is most abundant, cheap or green.

The concept is called 'demand dispatch', because it would activate and deactivate power demand — much as grid operators dynamically dispatch electricity generated by power plants today. In the future, power grids will probably use both the 'virtual storage' created by demand dispatch and the true energy storage from batteries. But demand dispatch could be the bigger player of the two, with smart use of existing appliances offering a smaller environmental footprint and slimmer price tag than batteries.

"You can create flexibility much more cheaply by controlling electrical loads," says Sean Meyn, director of the Florida Institute for Sustainable Energy in Gainesville. "You do have to install communications equipment, but after that there's nothing to wear out, nothing to replace."

Meyn is not alone in anticipating that demand dispatch — rather than batteries — will be the first line of defence for future grids (see 'Storage solutions'). A report published in May by the Massachusetts Institute of Technology Energy Initiative (go.nature.com/ogv7wa) said that controlling demand to maximize

consumption of solar power as it is generated is cheaper than storing excess power in batteries. "Load shifting to hours when [solar] generation is high should be the first resource to look at," notes the report.

This non-battery option is already coming online as researchers and commercial developers establish communication protocols to link energy-consuming equipment to grid operators, develop algorithms for optimizing their deployment and experiment with the business models required to finance the systems. But demand dispatch also faces concerns about its security and its financial benefits. A major test in the US Supreme Court this October could unleash the potential of demand dispatch, or leave it languishing on the sidelines to be eclipsed by battery storage.

REACTING TO DEMAND

Utilities companies around the world must keep spare power plants sitting idle to serve energy-demand spikes that may occur on just a few days each year. To keep the number of extra plants to a minimum, they use 'demand-side management' to help to control the need for power. This includes giving rebates for the

SÉBASTIEN THIBAUT

use of efficient appliances or implementing time-of-day pricing that incentivizes energy use when grid demand is low. Many utilities firms also practise 'demand response', in which they pay selected power users — usually factories and other large industrial and commercial operations — to cut their power consumption when the power grids are really strained.

Demand dispatch to provide virtual energy storage is an advanced form of demand response, the growth potential of which is limited by its disruptive impact on power users — shutting down a factory to save energy means lost revenue for the owners. Demand dispatch is about automating demand response and turning it into a more flexible, dynamic asset that customers hardly notice. The key is tapping into equipment that can be shut off for a few minutes or hours without inconveniencing users. Not all loads are a good fit, admits Mary Ann Piette, head of the Building Technology and Urban Systems Division at Lawrence Berkeley National Laboratory in California. Lighting is hard to switch off unobtrusively, for example.

But appropriate equipment is more common than you might think, she says. Electric water heaters, refrigerators and air-conditioning units can all be shut off for several hours at a time without affecting customers, thanks to thermal inertia, creating a window for virtual storage with little or no inconvenience.

Conventional demand response is exclusively about reducing load, but demand dispatch can often absorb electricity as well. Just as a battery can be charged and discharged, some electrical devices can be directed to pre-heat water, precool a space or bump up the pressure in a water line when strong winds and bright sunshine are providing surplus power. But aggregating many devices to create a substantial and reliable virtual-storage resource makes demand dispatch a more complex tool than batteries. "It's trickier," admits Robert Pratt, who leads the smart-grid programme at the Pacific Northwest National Laboratory in Richland, Washington. "But it may be cheaper."

Piette says that the system her group designed for commercial buildings costs US\$200–300 per kilowatt, considerably less than battery systems that can cost more than \$1,000 per kilowatt.

Meyn has studied the potential for controlling air-conditioning systems in commercial buildings to absorb the short-term minute-to-minute variability in solar power. He says that many commercial air-conditioning systems could keep supply and demand in balance by adjusting their fan speeds depending on the availability of solar power. His studies at a building on the University of Florida campus suggest that this would have no detectable effect on comfort, and would add no cost to running the equipment. The grid impact, meanwhile, looks large. "We can accommodate a grid powered 50% by renewable energy without the use

of batteries," says Meyn.

Entrepreneurial firms are already beginning to make money by doing what Meyn proposes. ENBALA Power Networks in Vancouver, Canada, which has been selling grid-balancing services since 2011, taps into large power loads that can be controlled variably, such as industrial water pumps, and adjusts them in response to signals from grid operators to use the available energy more efficiently.

Several demonstration projects completed in the past three years have tested demand dispatch as a means of smoothing out larger energy swings over several hours, such as the large shifts in wind-power output caused by changing weather systems. PowerShift Atlantic, a demonstration completed this year by a consortium of power-grid operators in three of Canada's maritime provinces, choreographed 17 megawatts of equipment at 1,400 homes and businesses to absorb bursts and troughs in the wind power that, at certain times, generates more than half of the region's electricity.

Controlling 20–30% of power demand would be enough to almost flatten out the variability in the region's power supply, says Tom Osterhus, chief executive of Integral Analytics, the company in Cincinnati, Ohio, that made the control software for PowerShift Atlantic. Michel Losier, the project's manager and director of customer and community engagement for utility firm NB Power in New Brunswick, Canada, says that on a typical winter's day, when the outside air temperature is often below freezing, water tanks and electric heating account for about 55% of New Brunswick's peak load — exactly the kinds of appliances that could be controlled to shape demand.

"We can accommodate a grid powered 50% by renewable energy without the use of batteries."

REMOTE CONTROL

Researchers and entrepreneurs face several technical hurdles to turn projects like PowerShift Atlantic into the reality for power-grid operation. Tapping the virtual-storage potential of millions of devices is a massive communications and control challenge. And, because many of those devices reside in homes, it means dealing with residential power customers.

Forging a reliable communications link between residential power devices and grid operators emerged as a challenge in both PowerShift Atlantic and a series of recent smart-grid projects funded by the US Department of Energy. Utilities often struggled to get various systems to communicate with each other. Other projects found that communication using smart meters, many of which are optimized to exchange data with the utility every 15 minutes, was too slow for centralized

dispatch of virtual storage.

As a result, operators are trying to bypass the smart meters and link to customers' home appliances using their broadband Internet connections. A standard protocol for Internet communications called OpenADR, which was developed by Piette's group, makes it easy for all the parts of the system to talk to each other. As the broad interconnection of consumer devices known as the Internet of Things progresses, costs will drop further and more loads will be able to participate, says Piette. "The future is rosy for virtual storage. Thanks to the Internet of Things, the cost of telemetry is getting cheaper."

But using the Internet comes with a caveat: it potentially leaves consumers and power grids vulnerable to cyberattacks and breaches in confidentiality. Pratt says that these risks will grow as connected appliances link to an ever-wider array of sites such as social media and online-shopping outfits. "If your refrigerator is not only connected to the grid but also running through the Wild Wild West of the Internet, that may not be good," he says. "It's an emerging concern that in the Internet of Things your devices will be so promiscuous in who they talk to that we're really opening up Pandora's box."

Pratt, Meyn and others are working on control schemes that could reduce the cybersecurity threat by relying on more distributed intelligence, with decisions made by the devices themselves rather than a central grid computer. Meyn is developing a distributed control scheme that he hopes will ultimately require no more than one digital handshake per day, rather than constant communication between devices and the grid.

Then again, Internet connectivity also offers powerful opportunities, and Internet companies such as Google are keen to take advantage. Google acquired home-automation firm Nest Labs, based in Palo Alto, California, for \$3.2 billion in 2014, and it is using it to build a virtual energy-storage system.

Nest's adaptive thermostats use sensors to guess when homes are unoccupied and then automatically put their heaters or air-conditioning units into a low-energy mode, making energy conservation easy. But they are also connected to the Internet, enabling the thermostat to factor local weather predictions into its behaviour. And, for homeowners who sign up to Nest's demand-dispatch programmes, connectivity creates a pool of potential virtual storage that utilities can access, through Nest, to dynamically adjust energy demand.

Nine utilities in the United States and Canada have signed up for Nest's demand-dispatch programmes, including Austin Energy in Texas, which had enrolled nearly 9,000 customers with smart thermostats by June. Together those homes can absorb or release up to 10.7 megawatts of power — a virtual storage capability that the utility expects to use 12–15 times per year to control demand spikes on hot days,

ALTERNATIVE ENERGY

Storage solutions



In pumped-storage hydropower stations, water is released to create energy when demand is high.

Grid-scale batteries face competition not only from 'virtual storage' systems that manipulate power demand, but also from other energy-storage technologies. Well-established alternatives include pumped storage hydropower, flywheels and compressed air.

Pumped-storage plants are the most affordable and proven means of large-scale energy storage, and they account for 97.5% of energy-storage capacity installed on global power grids, according to the US Department of Energy. Most plants charge up by pushing water uphill using surplus energy generated overnight (when consumption is low); during peak load times, the water flows downhill and spins a turbine to generate power. Utilities are now eyeing pumped storage as a means of managing growing flows of renewable energy.

For example, the Kauai Island Utility Cooperative (KIUC), which runs the power grid for the Hawaiian island of Kauai, plans to spend between US\$55 million and \$65 million to repurpose disused sugar-plantation reservoirs to create a 25-megawatt pumped-storage plant to help it to manage solar power. The KIUC estimates that storing energy from solar panels for use at night will be roughly 35% cheaper than running its oil-fired power plants. It will take time. The KIUC began studies in earnest in 2014, and says that it will be 2019 or later before the plant is built.

Flywheels, mechanical devices that are used to generate a power reserve, compete with batteries in applications that frequently require large power bursts for short periods, such as buffering power lines serving industrial mining equipment — heavy cycling conditions that degrade the capacity of batteries. Several recent flywheel projects target smaller power systems coping with high levels of intermittent renewable energy, using the flywheels to absorb and release power within microseconds to stabilize power demand in real time. In June, ABB in Zürich, Switzerland, installed a flywheel system on Alaska's Kodiak Island to help the local utility to cope with high levels of wind power. Darron Scott, president of the Kodiak Electric Association, says that the 2-megawatt flywheel replaced a 3-year-old lead-acid battery system that was wearing out faster than expected.

Another storage option is using electricity to compress air, which can then be converted back to electricity by using the air to spin a gas turbine. Early systems used underground caverns, including large installations in Alabama and in Germany. However, similar projects have been stymied by high equipment costs, energy losses and limited availability of appropriate geological formations. More-recent efforts focus on portable, distributed systems that store compressed air in steel or composite tanks. **P.F.**

when wholesale energy costs can rise from less than \$50 per megawatt-hour to more than \$1,000.

POWER PITFALLS

The biggest barrier to demand dispatch may be regulatory uncertainty over how it will be incorporated into the energy system, and how much grid operators should pay for it. There are positive signs for the market, as some states and regions forge ahead with rules that embrace demand dispatch, including PJM Interconnection, the US grid operator for 13 mid-Atlantic states and the District of Columbia.

Under US President Barack Obama, the Federal Energy Regulatory Commission (FERC) has enthusiastically embraced demand-side management. In 2011, FERC's Order 745 authorized wholesale power markets to pay companies that aggregate demand response the same price for managing megawatt-hours of demand as power-plant operators receive for generating energy. New York's Public Service Commission announced in June a mandate for all of the state's utilities to offer load-management programmes by mid-2016.

Demand response is also growing rapidly internationally. European countries are opening their energy markets to demand-response resources, and Japan and China are running pilot projects. Navigant Research, which specializes in clean-technology markets, predicts that annual global spending on demand response will grow from \$2 billion in 2015 to more than \$12.8 billion by 2023 as it expands beyond North America.

Despite these positive signs, there are still many energy markets in which regulators are taking a cautious approach, questioning whether the megawatt-hours of virtual storage that the technology promises will actually materialize when they are needed. This October, the US Supreme Court will decide whether to uphold a challenge — that compensation is a matter for state regulators — to FERC Order 745 by power generators. Advocates of demand dispatch, such as Meyn, say that the ruling is crucial for the technology's future.

Mei Shibata, co-founder of ThinkEco, a virtual-storage start-up in New York City, says that court challenges have already slowed growth in the United States by delaying the start of demonstration projects needed to convince energy regulators that the technology works. She predicts a big shakeout coming as venture-funded companies developing demand that can be dispatched run short of funds waiting for markets to mature. "The market will grow for residential and automated demand response, and convergence of demand response and the Internet of Things will happen, giving utilities better control over energy use," she says. "But how that happens isn't going to be a very pretty picture." ■

Peter Fairley is a freelance science writer based in Victoria, British Columbia.

IMAGEBROKER/ALAMY

BATTERIES

4 BIG QUESTIONS

The energy density of batteries will need to be substantially increased and their cost decreased if renewable energy is to replace fossil fuels. Here are four important questions.

BY KATHERINE BOURZAC

QUESTION

WHY IT MATTERS

STATUS QUO

SOLUTIONS

1

How do we make batteries less expensive?

The high price of batteries makes it more costly to integrate renewable energy — which can be intermittent and so needs to be stored — into the grid. It also means that the battery pack makes up the lion's share of the cost of electric vehicles.

The true cost of a battery is not only the cell cost, but also the packaging, installation, lifetime and storage capacity. Tesla Motors's 10-kilowatt-hour home storage system could cost about US\$7,000, once all installation costs have been factored in.

Companies can improve their packaging technology to lower costs without having to make changes to the fundamental chemistry of the battery. Jun Liu at Pacific Northwest National Laboratory in Washington, says that the cost needs to be reduced by half in the next few years.

2

How can we make batteries that store more energy?

Increasing the amount of energy a battery can store in a given weight or volume is the best way to bring down costs.

The gold standard for battery energy-storage capacity is lithium ion, which is around 250 watt-hours per kilogram.

Many battery electrode materials under development can store up to ten times more energy by weight than lithium-ion. On the anode side, these include silicon or pure lithium; on the cathode side alternatives include sulfur or air.

3

How can we make batteries safer?

If something goes wrong, such as overheating or a short circuit, batteries can catch fire. The Boeing 787 Dreamliner fleet was grounded after batteries caught fire on planes in 2013. No one was injured, but the incidents caused people to imagine a worst-case scenario.

Batteries on the market today are generally safe. But as battery packs get bigger, are more widely adopted and store more energy, risks will increase.

Researchers are coming up with protective coatings to prevent the growth of lithium-metal dendrites that cause short circuits, working on packing battery cells so that a fire in one cell will not spread to the next, and designing systems to sense problems and turn the battery off before a fire starts.

4

Are batteries enough for a renewable grid?

A full transition to renewable energy will require more battery capacity than currently exists in the world.

World production of lithium-ion batteries is about 35 gigawatt-hours per year. That will increase, but even with new electrode materials it is doubtful whether batteries can meet the world's electricity needs.

Non-electrochemical storage technologies are needed. Pumped hydropower can store a lot of energy, but it only works where there is water and mountains; other geographies are better suited to geothermal storage.

Katherine Bourzac is a science journalist based in San Francisco.

An ultrafast rechargeable aluminium-ion battery

Meng-Chang Lin^{1,2*}, Ming Gong^{1*}, Bingan Lu^{1,3*}, Yingpeng Wu^{1*}, Di-Yan Wang^{1,4,5}, Mingyun Guan¹, Michael Angell¹, Changxin Chen¹, Jiang Yang¹, Bing-Joe Hwang⁶ & Hongjie Dai¹

The development of new rechargeable battery systems could fuel various energy applications, from personal electronics to grid storage^{1,2}. Rechargeable aluminium-based batteries offer the possibilities of low cost and low flammability, together with three-electron-redox properties leading to high capacity³. However, research efforts over the past 30 years have encountered numerous problems, such as cathode material disintegration⁴, low cell discharge voltage (about 0.55 volts; ref. 5), capacitive behaviour without discharge voltage plateaus (1.1–0.2 volts⁶ or 1.8–0.8 volts⁷) and insufficient cycle life (less than 100 cycles) with rapid capacity decay (by 26–85 per cent over 100 cycles)^{4–7}. Here we present a rechargeable aluminium battery with high-rate capability that uses an aluminium metal anode and a three-dimensional graphitic-foam cathode. The battery operates through the electrochemical deposition and dissolution of aluminium at the anode, and intercalation/de-intercalation of chloroaluminate anions in the graphite, using a non-flammable ionic liquid electrolyte. The cell exhibits well-defined discharge voltage plateaus near 2 volts, a specific capacity of about 70 mA h g^{−1} and a Coulombic efficiency of approximately 98 per cent. The cathode was found to enable fast anion diffusion and intercalation, affording charging times of around one minute with a current density of ~4,000 mA g^{−1} (equivalent to ~3,000 W kg^{−1}), and to withstand more than 7,500 cycles without capacity decay.

Owing to the low-cost, low-flammability and three-electron redox properties of aluminium (Al), rechargeable Al-based batteries could in principle offer cost-effectiveness, high capacity and safety, which would

lead to a substantial advance in energy storage technology^{3,8}. However, research into rechargeable Al batteries over the past 30 years has failed to compete with research in other battery systems. This has been due to problems such as cathode material disintegration⁴, low cell discharge voltage (~0.55 V; ref. 5), capacitive behaviour without discharge voltage plateaus (1.1–0.2 V, or 1.8–0.8 V; refs 6 and 7, respectively), and insufficient cycle life (<100 cycles) with rapid capacity decay (by 26–85% over 100 cycles)^{4–7}. Here we report novel graphitic cathode materials that afford unprecedented discharge voltage profiles, cycling stabilities and rate capabilities for Al batteries.

We constructed Al/graphite cells (see diagram in Fig. 1a) in Swagelok or pouch cells, using an aluminium foil (thickness ~15–250 μm) anode, a graphitic cathode, and an ionic liquid electrolyte made from vacuum dried AlCl₃/1-ethyl-3-methylimidazolium chloride ([EMIm]Cl; see Methods, residual water ~500 p.p.m.). The cathode was made from either pyrolytic graphite (PG) foil (~17 μm) or a three-dimensional graphitic foam^{9,10}. Both the PG foil and the graphitic-foam materials exhibited typical graphite structure, with a sharp (002) X-ray diffraction (XRD) graphite peak at 2θ ≈ 26.55° (d spacing, 3.35 Å; Extended Data Fig. 1). The cell was first optimized in a Swagelok cell operating at 25 °C with a PG foil cathode. The optimal ratio of AlCl₃/[EMIm]Cl was found to be ~1.3–1.5 (Extended Data Fig. 2a), affording a specific discharging capacity of 60–66 mA h g^{−1} (based on graphitic cathode mass) with a Coulombic efficiency of 95–98%. Raman spectroscopy revealed that with an AlCl₃/[EMIm]Cl ratio of ~1.3, both AlCl₄[−] and Al₂Cl₇[−] anions were present (Extended Data Fig. 2b) at a ratio [AlCl₄[−]]/[Al₂Cl₇[−]] ≈ 2.33

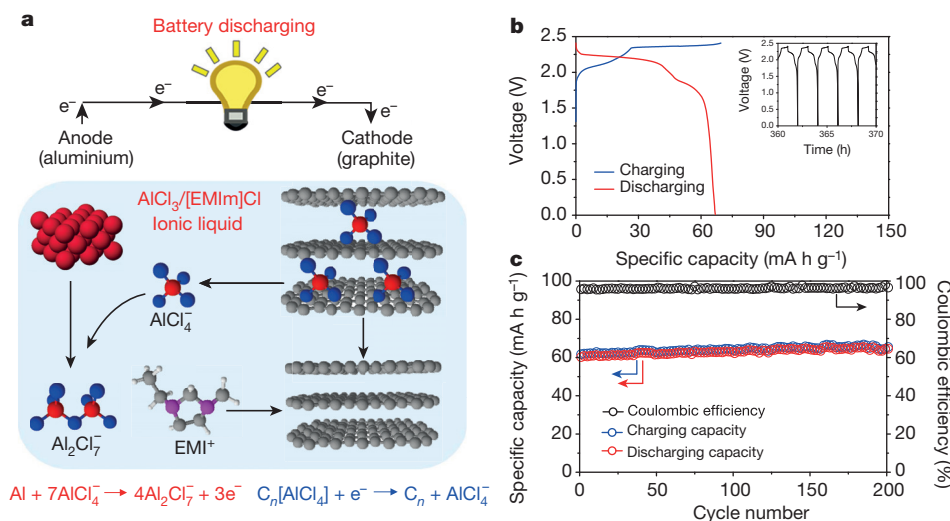


Figure 1 | Rechargeable Al/graphite cell. **a**, Schematic drawing of the Al/graphite cell during discharge, using the optimal composition of the AlCl₃/[EMIm]Cl ionic liquid electrolyte. On the anode side, metallic Al and AlCl₄[−] were transformed into Al₂Cl₇[−] during discharging, and the reverse reaction took place during charging. On the cathode side, predominantly AlCl₄[−] was

intercalated and de-intercalated between graphite layers during charge and discharge reactions, respectively. **b**, Galvanostatic charge and discharge curves of an Al/pyrolytic graphite (PG) Swagelok cell at a current density of 66 mA g^{−1}. Inset, charge and discharge cycles. **c**, Long-term stability test of an Al/PG cell at 66 mA g^{−1}.

¹Department of Chemistry, Stanford University, Stanford, California 94305, USA. ²Green Energy and Environment Research Laboratories, Industrial Technology Research Institute, Hsinchu 31040, Taiwan.

³School of Physics and Electronics, Hunan University, Changsha 410082, China. ⁴Department of Chemistry, National Taiwan Normal University, Taipei 11677, Taiwan. ⁵Institute of Atomic and Molecular Sciences, Academia Sinica, Taipei 10617, Taiwan. ⁶Department of Chemical Engineering, National Taiwan University of Science and Technology, Taipei 10607, Taiwan.

*These authors contributed equally to this work.

(ref. 11). The cathode specific discharging capacity was found to be independent of graphite mass (Extended Data Fig. 3), suggesting that the entirety of the graphite foil participated in the cathode reaction.

The Al/PG cell exhibited clear discharge voltage plateaus in the ranges 2.25–2.0 V and 1.9–1.5 V (Fig. 1b). The relatively high discharge voltage plateaus are unprecedented among all past Al-ion charge-storage systems^{4–7}. Similar cell operation was observed with the amount of electrolyte lowered to ~ 0.02 ml per mg of cathode material (Extended Data Fig. 4). Charge–discharge cycling at a current density of 66 mA g^{-1} (1 C charging rate) demonstrated the high stability of the Al/PG cell, which nearly perfectly maintained its specific capacity over >200 cycles with a $98.1 \pm 0.4\%$ Coulombic efficiency (Fig. 1c). This was consistent with the high reversibility of Al dissolution/deposition, with Coulombic efficiencies of 98.6–99.8% in ionic liquid electrolytes^{12–15}. No dendrite formation was observed on the Al electrode after cycling (Extended Data Fig. 5). To maintain a Coulombic efficiency $>96\%$, the cut-off voltage of the Al/PG cell (that is, the voltage at which charging was stopped) was set at 2.45 V, above which reduced efficiencies were observed (see Extended Data Fig. 6a), probably due to side reactions (especially above ~ 2.6 V) involving the electrolyte, as probed by cyclic voltammetry with a glassy carbon electrode against Al (Extended Data Fig. 6b).

We observed lowered Coulombic efficiency and cycling stability of the Al/graphite cell when using electrolytes with higher water contents, up to $\sim 7,500$ p.p.m. (Extended data Fig. 6c, d), accompanied by obvious H_2 gas evolution measured by gas chromatography (Extended Data Fig. 6e). This suggested side reactions triggered by the presence of residual water in the electrolyte, with H_2 evolution under reducing potential on the Al side during charging. Further lowering the water content

of the ionic liquid electrolyte could be important when maximizing the Coulombic efficiency of the Al/graphite cells.

The Al/PG cell showed limited rate capability with much lower specific capacity when charged and discharged at a rate higher than 1 C (Extended Data Fig. 7). It was determined that cathode reactions in the Al/PG cell involve intercalation and de-intercalation of relatively large chloroaluminate (Al_xCl_y^-) anions in the graphite (see below for XRD evidence of intercalation), and the rate capability is limited by slow diffusion of anions through the graphitic layers¹⁶. When PG was replaced by natural graphite, intercalation was evident during charging owing to dramatic expansion (~ 50 -fold) of the cathode into loosely stacked flakes visible to the naked eye (Extended Data Fig. 8a). In contrast, expansion of PG foil upon charging the Al/PG cell was not observable by eye (Extended Data Fig. 8b), despite the similar specific charging capacity of the two materials (Extended Data Fig. 8c). This superior structural integrity of PG over natural graphite during charging was attributed to the existence of covalent bonding between adjacent graphene sheets in PG¹⁷, which was not present in natural graphite. Using PG, which has an open, three-dimensionally-bound graphitic structure, we prevented excessive electrode expansion that would lead to electrode disintegration, while maintaining the efficient anion intercalation necessary for high performance.

Because high-rate and high-power batteries are highly desirable for applications such as electrical grid storage, the next step in the investigation was to develop a cathode material that would have reduced energetic barriers to intercalation during charging¹⁶. We investigated a flexible graphitic foam (Fig. 2a), which was made on a nickel foam template by chemical vapour deposition^{9,10} (see Methods), as a possible material for

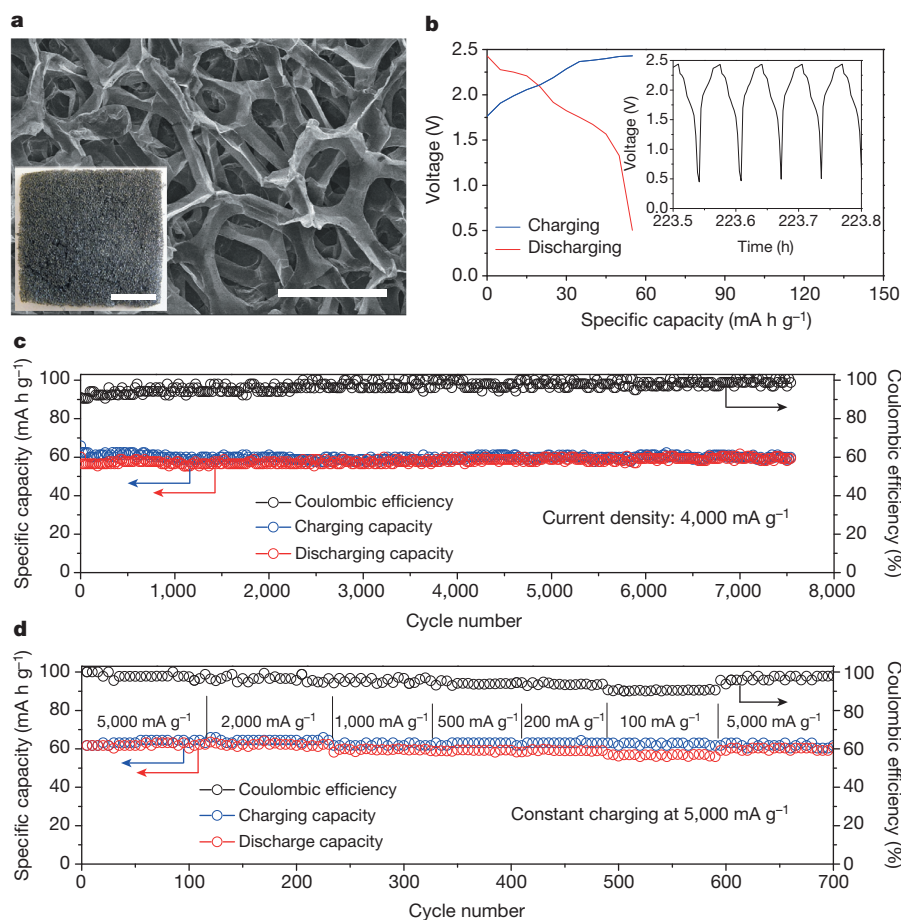


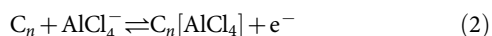
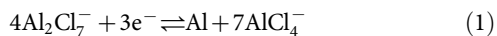
Figure 2 | An ultrafast and stable rechargeable Al/graphite cell. **a**, A scanning electron microscopy image showing a graphitic foam with an open frame structure; scale bar, 300 μm. Inset, photograph of graphitic foam; scale bar, 1 cm. **b**, Galvanostatic charge and discharge curves of an Al/graphitic-foam pouch cell at

a current density of $4,000 \text{ mA g}^{-1}$. **c**, Long-term stability test of an Al/graphitic-foam pouch cell over 7,500 charging and discharging cycles at a current density of $4,000 \text{ mA g}^{-1}$. **d**, An Al/graphitic-foam pouch cell charging at $5,000 \text{ mA g}^{-1}$ and discharging at current densities ranging from 100 to $5,000 \text{ mA g}^{-1}$.

ultrafast Al batteries. The graphite whiskers in the foam were 100 μm in width (Fig. 2a), with large spaces in between, which greatly decreased the diffusion length for the intercalating electrolyte anions and facilitated more rapid battery operation.

Remarkably, the Al/graphitic-foam cell (in a pouch cell configuration) could be charged and discharged at a current density up to $5,000 \text{ mA g}^{-1}$, about 75 times higher (that is, at a 75 C rate, <1 min charge/discharge time) than the Al/PG cell while maintaining a similar voltage profile and discharge capacity ($\sim 60 \text{ mA h g}^{-1}$) (Figs 1b and 2b). An impressive cycling stability with $\sim 100\%$ capacity retention was observed over 7,500 cycles with a Coulombic efficiency of $97 \pm 2.3\%$ (Fig. 2c). This is the first time an ultrafast Al-ion battery has been constructed with stability over thousands of cycles. The Al/graphitic-foam cell retained similar capacity and excellent cycling stability over a range of charge-discharge rates ($1,000$ – $6,000 \text{ mA g}^{-1}$) with 85 – 99% Coulombic efficiency (Extended Data Fig. 9a). It was also found that this cell could be rapidly charged (at $5,000 \text{ mA g}^{-1}$, in ~ 1 min) and gradually discharged (down to 100 mA g^{-1} , Fig. 2d and Extended Data Fig. 9b) over ~ 34 min while maintaining a high capacity ($\sim 60 \text{ mA h g}^{-1}$). Such a rapid charging/variable discharging rate could be appealing in many real-world applications.

We propose that simplified Al/graphite cell redox reactions during charging and discharging can be written as:



where n is the molar ratio of carbon atoms to intercalated anions in the graphite. The balanced AlCl_4^- and Al_2Cl_7^- concentrations in the electrolyte allowed for an optimal charging capacity at the cathode, with abundant AlCl_4^- for charging/intercalation in graphite (equation (2)), and sufficient Al_2Cl_7^- concentration for charging/electrodeposition at the anode (equation (1)).

Ex situ XRD measurement of graphite foil (Fig. 3a) confirmed graphite intercalation/de-intercalation by chloroaluminate anions during charging/discharging. The sharp pristine graphite foil (002) peak at $2\theta = 26.55^\circ$ (d spacing = 3.35 \AA) (Fig. 3a) vanished on charging to a specific capacity of $\sim 30 \text{ mA h g}^{-1}$, while two new peaks appeared at $\sim 28.25^\circ$ ($d \approx 3.15 \text{ \AA}$) and $\sim 23.56^\circ$ ($d \approx 3.77 \text{ \AA}$) (Fig. 3a), with peak intensities further increasing on fully charging to $\sim 62 \text{ mA h g}^{-1}$. The doublet XRD peak suggested highly strained graphene stacks formed on anion intercalation¹⁸. Analysis of the peak separation (see Methods) suggested a stage 4 graphite intercalation compound with an intercalant gallery height (spacing between adjacent graphitic host layers) of

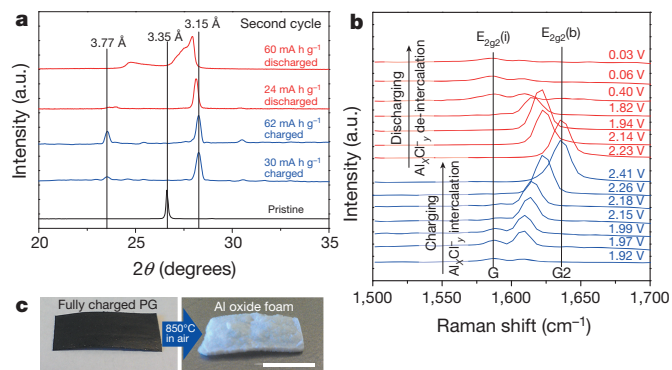


Figure 3 | Al/graphite cell reaction mechanisms. **a**, *Ex situ* X-ray diffraction patterns of PG in various charging and discharging states through the second cycle. **b**, *In situ* Raman spectra recorded for the PG cathode through a charge-discharge cycle, showing chloroaluminate anion intercalation/de-intercalation into graphite. **c**, After calcination of a fully charged (62 mA h g^{-1}) PG electrode at 850°C in air, the sample completely transformed into a white foam made of aluminium oxide. Scale bar, 1 cm.

$\sim 5.7 \text{ \AA}$, indicating that the AlCl_4^- anions (size $\sim 5.28 \text{ \AA}$; ref. 19) were intercalated between graphene layers in a distorted state. Full discharging led to the recovery of the graphite peak but with a broad shoulder (Fig. 3a), probably caused by irreversible changes in the stacking between the graphene layers or a small amount of trapped species.

In situ Raman spectroscopy was also performed to probe chloroaluminate anion intercalation/de-intercalation from graphite during cell charge/discharge (Fig. 3b). The graphite G band ($\sim 1,584 \text{ cm}^{-1}$) diminished and split into a doublet ($1,587 \text{ cm}^{-1}$ for the $\text{E}_{2g2}(i)$ mode and $\sim 1,608 \text{ cm}^{-1}$ for the $\text{E}_{2g2}(b)$ mode) upon anion intercalation (Fig. 3b)²⁰, and then evolved into a sharp new peak ($\sim 1,636 \text{ cm}^{-1}$, the G2 band of the $\text{E}_{2g2}(b)$ mode, spectrum 2.41 V, Fig. 3b) once fully charged. The spectral changes were then reversed upon discharging (Fig. 3b), as the typical graphite Raman G band (1584 cm^{-1}) was recovered when fully discharged (spectrum 0.03 V, Fig. 3b). Similar Raman spectra and XRD data were obtained with a graphitic-foam cathode (Extended Data Fig. 10a, b). Interestingly, calcination of a fully charged PG foil at 850°C in air (Fig. 3c) yielded a white aluminium oxide foam (Extended Data Fig. 10c), confirming the intercalation of chloroaluminate anions into the carbon network, which had been evidently removed oxidatively.

Lastly, X-ray photoelectron spectra (XPS) and Auger electron spectroscopy (AES) were performed to probe the chemical nature of the intercalated species in our graphitic cathodes (see Methods for details). To minimize the amount of trapped electrolyte, graphitic foam was used and the electrode was thoroughly washed with anhydrous methanol. XPS revealed that upon charging pristine graphite, the 284.8 eV C 1s peak developed a shoulder at higher energy ($\sim 285.9 \text{ eV}$, Fig. 4a), confirming electrochemical oxidation of graphitic carbon by intercalation of AlCl_4^- anions (equation (2)). Chloroaluminate intercalation was evident from the appearance of Al 2p and Cl 2p peaks (Fig. 4b, c). Upon

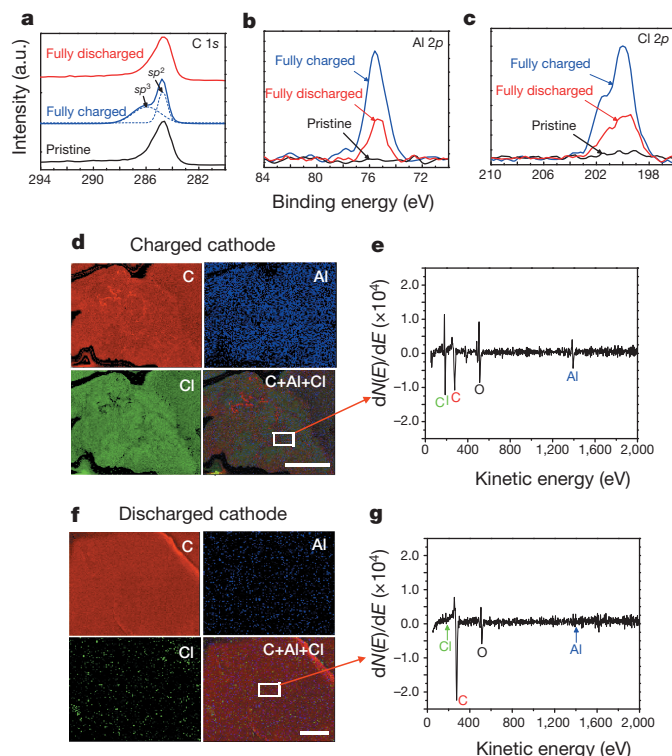


Figure 4 | Chemical probing of a graphitic cathode by XPS and AES. **a**, XPS data of the C 1s peak of a graphitic-foam electrode: pristine, fully charged and fully discharged. **b**, **c**, XPS data of Al 2p and Cl 2p peaks observed with a graphitic-foam electrode: pristine, fully charged and fully discharged. **d**–**g**, AES mapping images for C, Al and Cl (**d**, **f**), and the AES spectrum of the boxed regions (**e**, **g**) obtained with a fully charged graphitic-foam sample (**d**, **e**) and a fully discharged graphitic-foam sample (**f**, **g**). Scale bars: **d**, 25 μm ; **f**, 10 μm .

discharging, the C 1s XPS spectrum of the cathode reverted to that of the pristine graphite due to anion de-intercalation and carbon reduction (Fig. 4a). Also, a substantial reduction in the Al 2p and Cl 2p signals was recorded over the graphite sample (see Fig. 4b, c). The remaining Al and Cl signals observed were attributed to trapped/adsorbed species in the graphite sample, which was probed by XPS over a large area. Furthermore, high spatial resolution AES elemental mapping of a single graphite whisker in the fully charged graphitic foam clearly revealed Al and Cl Auger signals uniformly distributed over the whisker (Fig. 4d, e), again confirming chloroaluminate anion intercalation. When fully discharged, AES mapping revealed anion de-intercalation from graphite with much lower Al and Cl Auger signals observed (Fig. 4f, g). These spectroscopic results clearly revealed chloroaluminate ion intercalation/de-intercalation in the graphite redox reactions involved in our rechargeable Al cell.

The Al battery pouch cell is mechanically bendable and foldable (Supplementary Video 1) owing to the flexibility of the electrode and separator materials. Further, we drilled through Al battery pouch cells during battery operation and observed no safety hazard, owing to the lack of flammability of the ionic liquid electrolyte in air (see Supplementary Video 2).

We have developed a new Al-ion battery using novel graphitic cathode materials with a stable cycling life up to 7,500 charge/discharge cycles without decay at ultrahigh current densities. The present Al/graphite battery can afford an energy density of $\sim 40 \text{ W h kg}^{-1}$ (comparable to lead-acid and Ni-MH batteries, with room for improvement by optimizing the graphitic electrodes and by developing other novel cathode materials) and a high power density, up to $3,000 \text{ W kg}^{-1}$ (similar to supercapacitors). We note that the energy/power densities were calculated on the basis of the measured $\sim 65 \text{ mA h g}^{-1}$ cathode capacity and the mass of active materials in electrodes and electrolyte. Such rechargeable Al ion batteries have the potential to be cost effective and safe, and to have high power density.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 12 March 2014; accepted 6 February 2015.

Published online 6 April 2015.

1. Yang, Z. *et al.* Electrochemical energy storage for green grid. *Chem. Rev.* **111**, 3577–3613 (2011).
2. Huskinson, B. *et al.* A metal-free organic-inorganic aqueous flow battery. *Nature* **505**, 195–198 (2014).
3. Li, Q. & Bjerrum, N. J. Aluminum as anode for energy storage and conversion: a review. *J. Power Sources* **110**, 1–10 (2002).
4. Gifford, P. R. & Palmisano, J. B. An aluminum/chlorine rechargeable cell employing a room temperature molten salt electrolyte. *J. Electrochem. Soc.* **135**, 650–654 (1988).
5. Jayaprakash, N., Das, S. K. & Archer, L. A. The rechargeable aluminum-ion battery. *Chem. Commun.* **47**, 12610–12612 (2011).
6. Rani, J. V., Kanakaiah, V., Dadmal, T., Rao, M. S. & Bhavanarushi, S. Fluorinated natural graphite cathode for rechargeable ionic liquid based aluminum-ion battery. *J. Electrochem. Soc.* **160**, A1781–A1784 (2013).

7. Hudak, N. S. Chloroaluminate-doped conducting polymers as positive electrodes in rechargeable aluminum batteries. *J. Phys. Chem. C* **118**, 5203–5215 (2014).
8. Armand, M. & Tarascon, J. M. Building better batteries. *Nature* **451**, 652–657 (2008).
9. Yu, X., Lu, B. & Xu, Z. Super long-life supercapacitors based on the construction of nanohoneycomb-like strongly coupled CoMoO_4 -3D graphene hybrid electrodes. *Adv. Mater.* **26**, 1044–1051 (2014).
10. Chen, Z. *et al.* Three-dimensional flexible and conductive interconnected graphene networks grown by chemical vapour deposition. *Nature Mater.* **10**, 424–428 (2011).
11. Wasserscheid, P. & Keim, W. Ionic liquids—new “solutions” for transition metal catalysis. *Angew. Chem. Int. Edn* **39**, 3772–3789 (2000).
12. Auburn, J. J. & Barberio, Y. L. An ambient temperature secondary aluminum electrode: its cycling rates and its cycling efficiencies. *J. Electrochem. Soc.* **132**, 598–601 (1985).
13. Wilkes, J. S., Levisky, J. A., Wilson, R. A. & Hussey, C. L. Dialkylimidazolium chloroaluminate melts: a new class of room-temperature ionic liquids for electrochemistry, spectroscopy and synthesis. *Inorg. Chem.* **21**, 1263–1264 (1982).
14. Lai, P. K. & Skylas-Kazacos, M. Electrodeposition of aluminium in aluminium chloride/1-methyl-3-ethylimidazolium chloride. *J. Electroanal. Chem. Interfacial Electrochem.* **248**, 431–440 (1988).
15. Jiang, T., Chollier Brym, M. J., Dubé, G., Lasia, A. & Brisard, G. M. Electrodeposition of aluminium from ionic liquids: Part I—electrodeposition and surface morphology of aluminium from aluminium chloride (AlCl_3)–1-ethyl-3-methylimidazolium chloride ($[\text{EMIm}]\text{Cl}$) ionic liquids. *Surf. Coat. Tech.* **201**, 1–9 (2006).
16. Borg, R. J. & Dienes, G. J. *An Introduction to Solid State Diffusion* (Academic, 1988).
17. Zhu, Y.-J., Hansen, T. A., Ammermann, S., McBride, J. D. & Beebe, T. P. Nanometer-size monolayer and multilayer molecule corrals on HOPG: a depth-resolved mechanistic study by STM. *J. Phys. Chem. B* **105**, 7632–7638 (2001).
18. Schmuelling, G. *et al.* X-ray diffraction studies of the electrochemical intercalation of bis(trifluoromethanesulfonyl)imide anions into graphite for dual-ion cells. *J. Power Sources* **239**, 563–571 (2013).
19. Takahashi, S., Koura, N., Kohara, S., Saboungi, M. L. & Curtiss, L. A. Technological and scientific issues of room-temperature molten salts. *Plasma Ion* **2**, 91–105 (1999).
20. Hardwick, L. J. *et al.* An *in situ* Raman study of the intercalation of supercapacitor-type electrolyte into microcrystalline graphite. *Electrochim. Acta* **52**, 675–680 (2006).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank M. D. Fayer for discussions. We also thank Y. Cui's group for use of an argon-filled glove box and a vacuum oven. M.-C.L. thanks the Bureau of Energy, Ministry of Economic Affairs, Taiwan, for supporting international cooperation between Stanford University and ITRI. B.L. acknowledges support from the National Natural Science Foundation of China (grant no. 21303046), the China Scholarship Council (no. 201308430178), and the Hunan University Fund for Multidisciplinary Developing (no. 531107040762). We also acknowledge support from the US Department of Energy for novel carbon materials development and electrical characterization work (DOE DE-SC0008684), Stanford GCEP, the Precourt Institute of Energy, and the Global Networking Talent 3.0 plan (NTUST 104DI005) from the Ministry of Education of Taiwan.

Author Contributions M.-C.L., M.G., B.L. and Y.W. contributed equally to this work. M.-C.L. and H.D. conceived the idea for the project. B.L. prepared the graphitic foam. M.-C.L., M.G., B.L., Y.W., D.-Y.W., M.A. and M. Guan performed electrochemical experiments. M.-C.L., C.C. and J.Y. conducted *in situ* Raman spectroscopy measurements. M.-C.L., M.G., B.L. and Y.W. performed *ex situ* X-ray diffraction measurements. M.G., M.-C.L., B.L. and Y.W. performed X-ray photoelectron spectroscopy and Auger electron spectroscopy measurements. M.-C.L., M.G., B.L., Y.W., D.-Y.W., M.A., B.-J.H. and H.D. discussed the results, analysed the data and drafted the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to H.D. (hdoi@stanford.edu).

METHODS

Preparation of ionic liquid electrolytes. A room temperature ionic liquid electrolyte was made by mixing 1-ethyl-3-methylimidazolium chloride ([EMIm]Cl, 97%, Acros Chemicals) and anhydrous aluminium chloride (AlCl_3 , 99.999%, Sigma Aldrich). [EMIm]Cl was baked at 130 °C under vacuum for 16–32 h to remove residual water. ([EMIm] Al_2Cl_7) ionic liquid electrolytes were prepared in an argon-atmosphere glove box (both [EMIm]Cl and AlCl_3 are highly hygroscopic) by mixing anhydrous AlCl_3 with [EMIm]Cl, and the resulting light-yellow, transparent liquid was stirred at room temperature for 10 min. The mole ratio of AlCl_3 to [EMIm]Cl was varied from 1.1 to 1.8. The water content of the ionic liquid was determined (500–700 p.p.m.) using a coulometric Karl Fischer titrator, DL 39 (Mettler Toledo). The predominant anions in basic melts (AlCl_3 /[EMIm]Cl mole ratio <1) are Cl^- and AlCl_4^- , while in acidic melts (AlCl_3 /[EMIm]Cl mole ratio >1) chloroaluminate anions such as Al_2Cl_7^- , $\text{Al}_3\text{Cl}_{10}^-$, and $\text{Al}_4\text{Cl}_{13}^-$ are formed¹¹. The ratio of anions to cations in the AlCl_3 /[EMIm]Cl electrolyte was determined using a glass fibre filter paper (Whatman GF/D) loaded with a 4–8 μm Au-coated SiO_2 beads²¹ in a cuvette cell (0.35 ml, Starna Cells) with random orientation quartz windows. Then, in the glove box, the cuvette cell was filled with AlCl_3 /[EMIm]Cl = 1.3 (by mole). Raman spectra (200–650 cm^{-1}) were obtained using a 785-nm laser with 2 cm^{-1} resolution. Raman data were collected from the surface of the Au-coated SiO_2 bead so as to benefit from surface enhanced Raman^{21,22} (Extended Data Fig. 2b).

Preparation of graphitic foam. Nickel (Ni) foams (Alantum Advanced Technology Materials, Shenyang, China), were used as 3D scaffold templates for the CVD growth of graphitic foam, following the process reported previously^{9,10}. The Ni foams were heated to 1,000 °C in a horizontal tube furnace (Lindberg Blue M, TF55030C) under Ar (500 standard cubic centimetres per minute or s.c.c.m.) and H_2 (200 s.c.c.m.) and annealed for 10 min to clean their surfaces and to eliminate a thin surface oxide layer. Then, methane (CH_4) was introduced into the reaction tube at ambient pressure at a flow rate of 10 s.c.c.m., corresponding to a concentration of 1.4 vol.% in the total gas flow. After 10 min of reaction gas mixture flow, the samples were rapidly cooled to room temperature at a rate of 300 °C min^{-1} under Ar (500 s.c.c.m.) and H_2 (200 s.c.c.m.). The Ni foams covered with graphite were drop-coated with a poly(methyl methacrylate) (PMMA) solution (4.5% in ethyl acetate), and then baked at 110 °C for 0.5 h. The PMMA/graphene/Ni foam structure was obtained after solidification. Afterwards, these samples were put into a 3 M HCl solution for 3 h to completely dissolve the Ni foam to obtain the PMMA/graphite at 80 °C. Finally, the pure graphitic foam was obtained by removing PMMA in hot acetone at 55 °C and annealing in NH_3 (80 s.c.c.m.) at 600 °C for 2 h, and then annealing in air at 450 °C for 2 h. The microstructure of the graphitic foam was examined by SEM analysis using a FEI XL30 Sirion scanning electron microscope (Fig. 2a in the main text).

Preparation of glassy carbon. Glassy carbon (GC) was used as the current collector in the Swagelok-type cell. 72 g phenol (Sigma-Aldrich) and 4.5 ml ammonium hydroxide (30%, Fisher Scientific) were dissolved in 100 ml formaldehyde solution (37%, Fisher Scientific) under reflux while stirring rapidly. The solution was stirred at 90 °C until the solution turned a milk-white colour. Rotary evaporation was used to remove the water and get the phenolic resin. The phenolic resin was solidified at 100 °C in a mould (1/2-inch glass tube), and then carbonized at 850 °C under an Ar atmosphere for four hours to obtain the GC rod. The resulting GC rod contributed negligible capacity to the cathode (Extended Data Fig. 6b).

Electrochemical measurements. Prior to assembling the Al/graphite cell in the glove box, all components were heated under vacuum at 60 °C for more than 12 h to remove residual water. All electrochemical tests were performed at 25 ± 1 °C. A Swagelok-type cell (1/2 inch diameter) was constructed using a ~4 mg PG foil (0.017 mm, Suzhou Dasen Electronics Materials) cathode and a 90 mg Al foil (0.25 mm, Alfa Aesar) anode. A 1/2 inch GC rod (10 mm) was used as the current collector for the PG cathode, and a 1/2 inch graphite rod (10 mm) was used for the Al anode. Six layers of 1/2 inch glass fibre filter paper (Whatman 934-AH) were placed between the anode and cathode. Then, ~1.0 ml of ionic liquid electrolyte (prepared with AlCl_3 /[EMIm]Cl mole ratios of 1.1, 1.3, 1.5 and 1.8) was injected and the cell sealed. The Al/PG cell was then charged (to 2.45 V) and discharged (to 0.01 V) at a current density of 66 mA g^{-1} with a MTI battery analyser (BST8-WA) to identify the ideal AlCl_3 /[EMIm]Cl mole ratio (Extended Data Fig. 2a). To investigate the Coulombic efficiency of the Al/PG cell in AlCl_3 /[EMIm]Cl \approx 1.3 (by mole) electrolyte, the cell was charged to 2.45, 2.50, 2.55 and 2.60 V, respectively, and discharged to 0.4 V at a current density of 66 mA g^{-1} (Extended Data Fig. 6a). For long-term cycling stability tests, an Al/PG cell using electrolyte AlCl_3 /[EMIm]Cl \approx 1.3 by mole was charged/discharged at a current density of 66 mA g^{-1} (Fig. 1b, c in the main text). To study the rate capability of the Al/PG cell, the current densities were varied from 66 to 264 mA g^{-1} (Extended Data Fig. 7). Note that we lowered the electrolyte amount to ~0.02 ml per mg of cathode material and observed similar cell operation (Extended Data Fig. 4). Further decrease in the electrolyte ratio is possible through battery engineering.

PG foil was synthesized by pyrolysis of polyimide at high temperature, in which some covalent bonding is inevitably generated due to imperfections. Natural graphite foil was produced by compressing expanded graphite flakes, leading to stacking of natural graphite flakes by Van der Waals bonding between them. Similar battery characteristics were observed with PG and graphite foil electrodes, indicating that the battery behaviour was derived from the graphitic property of the electrodes (Extended Data Fig. 8c). However, since the natural graphite foils are synthesized by compressing expanded natural graphite powders without the covalent linkage between them, these foils suffered from drastic electrode expansion obvious to the naked eye, whereas pyrolytic graphite foils showed no obvious electrode expansion due to covalency (Extended Data Fig. 8a, b).

Pouch cells were assembled in the glove box using a graphitic-foam (~3 mg) cathode and an Al foil (~70 mg) anode, which were separated by two layers of glass fibre filter paper to prevent shorting. Polymer (0.1 mm \times 4 mm \times 5 mm) coated Ni foils (0.09 mm \times 3 mm \times 60 mm in size; MTI corporation) were used as current collectors for both anode and cathode. The electrolyte (~2 ml prepared using AlCl_3 /[EMIm]Cl = 1.3 by mole) was injected and the cell was closed using a heat sealer. The cell was removed from the glove box for long-term cycling stability tests, in which the cell was charged/discharged at a current density of 4,000 mA g^{-1} (Fig. 2b, c). To determine the rate capability and fast-charge/slow-discharge behaviours of the Al/graphitic-foam cell, various current densities from 100 to 5,000 mA g^{-1} were used (Extended Data Fig. 9 and Fig. 2d). The pouch cell was charged to 2.42 V and discharged to a cut-off voltage of 0.5 V to prevent the dissolution reaction of Ni foil in the ionic liquid electrolyte.

Cyclic voltammetry measurements were performed using a potentiostat/galvanostat model CHI 760D (CH Instruments) in either three-electrode or two-electrode mode. The working electrode was an Al foil or a PG foil, the auxiliary electrode consisted of an Al foil, and an Al foil was used as the reference electrode. Copper tape (3M) was attached to these electrodes as the current collector. The copper tape was covered by poly-tetrafluoroethylene (PTFE) tape to prevent contact with the ionic liquid electrolyte and the part of the copper tape covered by PTFE was not immersed in the ionic liquid electrolyte. This prevented corrosion of the copper tape during cyclic voltammetry measurements. All three electrodes were placed in a plastic (1.5 ml) cuvette cell (containing electrolyte AlCl_3 /[EMIm]Cl = 1.3 by mole) in the glove box, and then sealed with a rubber cap using a clamp. The scanning voltage range was set from -1.0 to 1.0 V (versus Al) for Al foil and 0 to 2.5 V (versus Al) for graphitic material, and the scan rate was 10 mV s^{-1} (Extended Data Fig. 10d). To investigate the working voltage range of the electrolyte without involving cathode intercalation, two-electrode measurement was performed by using a GC rod cathode against an Al anode in a Swagelok cell in AlCl_3 /[EMIm]Cl (~1.3 by mole) electrolyte. The scanning voltage range was set from 0 to 2.9 V at a scan rate of 10 mV s^{-1} (Extended Data Fig. 6b).

We investigated the Al ion cell operation mechanism and electrode reactions in the ionic liquid electrolyte, using the optimal mole ratio of AlCl_3 /[EMIm]Cl = 1.3. Using CV (Extended Data Fig. 10d), a reduction wave from -1.0 to -0.08 V (versus Al) and an oxidation wave from -0.08 to 0.80 V (versus Al) for the anode were observed (Extended Data Fig. 10d, left plot), corresponding to Al reduction/electrodeposition and oxidation/dissolution^{13,15,23–25} during charging and discharging, respectively. This was consistent with Al redox electrochemistry in chloroaluminate ionic liquids^{13,15,23–25} via equation (1) in the main text, and consistent with our Raman measurements, which showed both AlCl_4^- and Al_2Cl_7^- in the electrolyte (Extended Data Fig. 2b). On the graphitic cathode side, an oxidation wave of 1.83 to 2.50 V (versus Al) and a reduction wave of 1.16 to 2.36 V (versus Al) were observed (Extended Data Fig. 10d, right plot) and attributed to graphite oxidation and reduction through intercalation and de-intercalation of anions (predominantly AlCl_4^- due to its smaller size), respectively. The oxidation voltage range of 1.83 to 2.50 V (versus Al, Extended Data Fig. 10d, right plot) was close to the anodic voltage range (1.8 to 2.2 V versus Al) of a previously reported dual-graphite cell²⁶ attributed to AlCl_4^- intercalation in graphite. The reduction wave range of 1.16 to 2.36 V (versus Al) was assigned to the AlCl_4^- de-intercalation²⁶. The nature of the shoulder in the reduction curve of graphite ranging from 2.36 to 1.9 V (Extended Data Fig. 10d, right plot) and a higher discharge plateau (2.25 to 2.0 V) of an Al/PG cell upon charging (Fig. 1b in the main text) remained unclear, but could be due to different stages of anion-graphite intercalation²⁷.

XRD and Raman studies of graphite cathodes during charge and discharge. For *ex situ* X-ray diffraction (XRD) study, an Al/PG cell (in a Swagelok configuration) was charged and discharged at a constant current density of 66 mA g^{-1} . The reactions were stopped after 30 mA h g^{-1} charged, fully charged (62 mA h g^{-1}) and 40 mA h g^{-1} discharged after charge/discharge capacities were in a stable state. Fully charged (62 mA h g^{-1}) graphitic foam was also prepared. After either the charge or the discharge reaction, the graphitic cathode was removed from the cell in the glove box. To avoid reaction between the cathode and air/moisture in the ambient atmosphere, the cathode was placed onto a glass slide and then wrapped in a Scotch tape.

The wrapped samples were immediately removed from the glove box for *ex situ* XRD measurements, which were performed on a PANalytical X'Pert instrument (Fig. 3a in the main text and Extended Data Fig. 10b).

The periodic repeat distance (I_C), the intercalant gallery height (d_i) and the gallery expansion (Δd)^{28,29} can be calculated using

$$I_C = (d_i + 3.35 \text{ \AA}) \times (n - 1) = (\Delta d + 3.35 \text{ \AA}) \times n = l \times d_{\text{obs}} \quad (3)$$

where l is the index of ($00l$) planes oriented in the stacking direction and d_{obs} is the observed value of the spacing between two adjacent planes^{18,28,29}. The d spacing of graphite is 3.35 Å. During the charging/anion-intercalation process, the graphite (002) peak completely vanished and two new peaks arose. The intensity pattern is commonly found for a stage n graphite intercalation compound (GIC), where the most dominant peak is the ($00n + 1$) and the second most dominant peak is the ($00n + 2$)^{18,28,29}. Based on our experimental data, by increasing the charging state from 48–60% charged (30 mA h g^{-1}) to the fully charged state (62 mA h g^{-1}), the distance between the ($00n + 1$) and ($00n + 2$) peaks gradually increased, as more Al_xCl_y^- anions intercalated. The d spacing values of ($00n + 1$) and ($00n + 2$) peaks (that is, $d_{(n+1)}$ and $d_{(n+2)}$, respectively) were calculated from XRD data (for example, Fig. 3a). By determining the ratio of the $d_{(n+2)}/d_{(n+1)}$ peak position and correlating these to the ratios of stage pure GICs (that is, ideal cases), the most dominant stage phase of the observed GIC can be assigned^{28,29}. After assigning the ($00l$) indices, we calculated the intercalant gallery height (d_i) through equation (3).

For simultaneous *in situ* Raman and galvanostatic charge/discharge reaction measurements, a cuvette cell (0.35 ml, Starna Cells) with random orientation quartz windows was used. An aluminium foil and a graphitic material (PG or graphitic foam) were used as the anode and cathode, respectively. The electrolyte was mixed $\text{AlCl}_3/[\text{EMIm}]\text{Cl} = 1.3$ (by mole). The electrochemical cell was assembled in the glove box following the process mentioned above. Raman spectra were obtained ($1,500\text{--}1,700 \text{ cm}^{-1}$) using a HeNe laser (633 nm) with 2 cm^{-1} resolution. The spectral data were collected after a few successive charge/discharge scans between 2.45 and 0.01 V at a current density of 66 mA g^{-1} (PG) (Fig. 3b in the main text) or $1,000 \text{ mA g}^{-1}$ (graphitic foam) (Extended Data Fig. 10a).

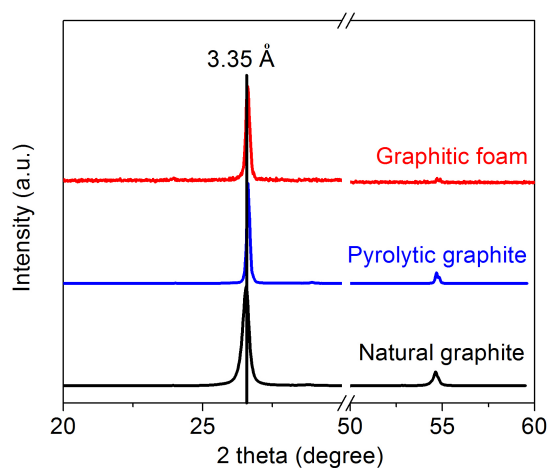
XPS and AES measurements. Al/graphitic-foam cells were fully charged/discharged at a current density of $4,000 \text{ mA g}^{-1}$. Then, the Al/graphitic-foam cells were transferred to the glove box for preparation for XPS and AES analysis. Fully charged/

discharged graphitic foams were collected from the pouch cell and washed with anhydrous methanol to remove the residual $\text{AlCl}_3/\text{EMIC}$ ionic liquid electrolyte. The as-rinsed graphitic foams were attached to a Si wafer and baked at 90°C for 10 min to remove residual methanol. The samples were sealed in a plastic pouch to avoid contamination by reaction with moisture and oxygen before XPS and AES characterization. Auger electron spectra were taken by a PHI 700 Scanning Auger Nanoprobe operating at 10 kV and 10 nA. XPS spectra were collected on a PHI VersaProbe Scanning XPS Microprobe (Fig. 4 in the main text).

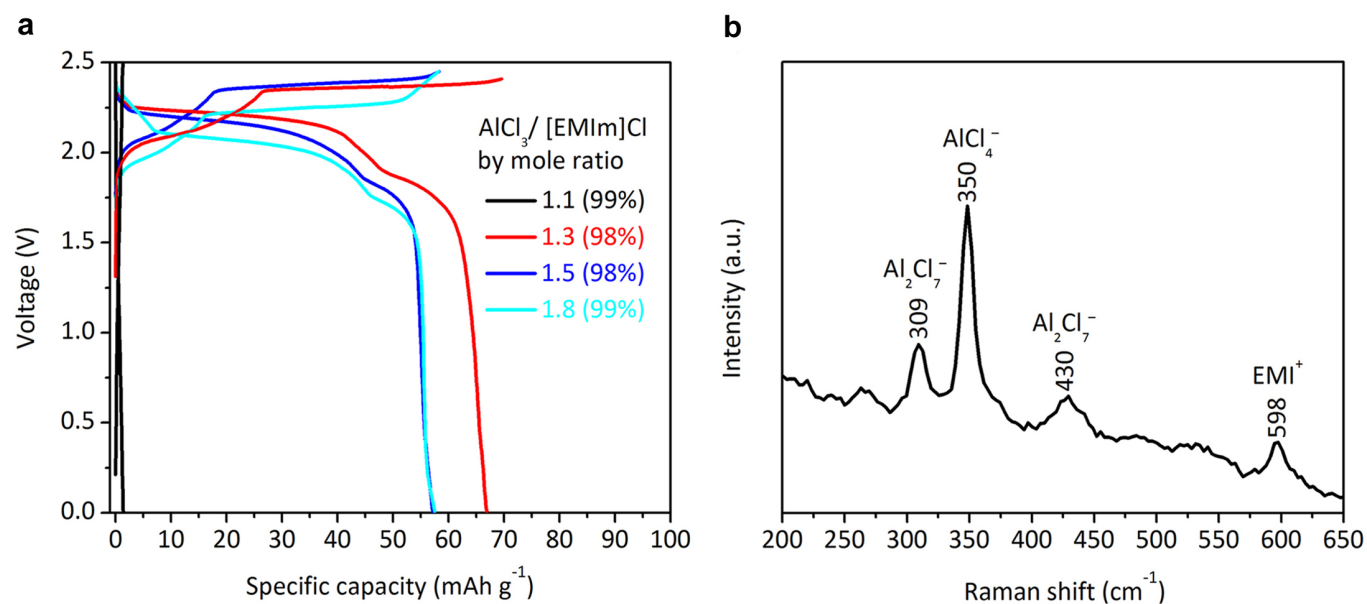
TGA measurements. Fully charged PG cathodes were washed with methanol for 24 h to remove the residual $\text{AlCl}_3/\text{EMIC}$ ionic liquid electrolyte. The as-washed PG samples were calcined at 850°C for 3 h in air. The as-calcined samples (white foam) were collected, weighed, and analysed by SEM-EDX to study the chemical composition (Extended Data Fig. 10c). SEM and SEM-EDX analyses were performed using an FEI XL30 Sirion scanning electron microscope.

Sample size. No statistical methods were used to predetermine sample size.

- Zhang, B. *et al.* Plasmonic micro-beads for fluorescence enhanced, multiplexed protein detection with flow cytometry. *Chem. Sci.* **5**, 4070–4075 (2014).
- Tabakman, S. M., Chen, Z., Casalongue, H. S., Wang, H. & Dai, H. A new approach to solution-phase gold seeding for SERS substrates. *Small* **7**, 499–505 (2011).
- Lee, J. J., Bae, I. T., Scherson, D. A., Miller, B. & Wheeler, K. A. Underpotential deposition of aluminum and alloy formation on polycrystalline gold electrodes from $\text{AlCl}_3/\text{EMIC}$ room-temperature molten salts. *J. Electrochem. Soc.* **147**, 562–566 (2000).
- Pan, S.-J., Tsai, W.-T., Chang, J.-K. & Sun, I. W. Co-deposition of Al–Zn on AZ91D magnesium alloy in AlCl_3 –1-ethyl-3-methylimidazolium chloride ionic liquid. *Electrochim. Acta* **55**, 2158–2162 (2010).
- Endres, F., MacFarlane, D. & Abbott, A. *Electrodeposition from Ionic Liquids* (Wiley & Sons, 2008).
- Carlin, R. T., De Long, H. C., Fuller, J. & Trulove, P. C. Dual intercalating molten electrolyte batteries. *J. Electrochem. Soc.* **141**, L73–L76 (1994).
- Bao, W. *et al.* Approaching the limits of transparency and conductivity in graphitic materials through lithium intercalation. *Nature Commun.* **5**, 4224 (2014).
- Zhang, X., Sukpirom, N. & Lerner, M. M. Graphite intercalation of bis(trifluoromethanesulfonyl) imide and other anions with perfluoroalkanesulfonyl substituents. *Mater. Res. Bull.* **34**, 363–372 (1999).
- Özmen-Monkul, B. & Lerner, M. M. The first graphite intercalation compounds containing tris(pentafluoroethyl)trifluorophosphate. *Carbon* **48**, 3205–3210 (2010).

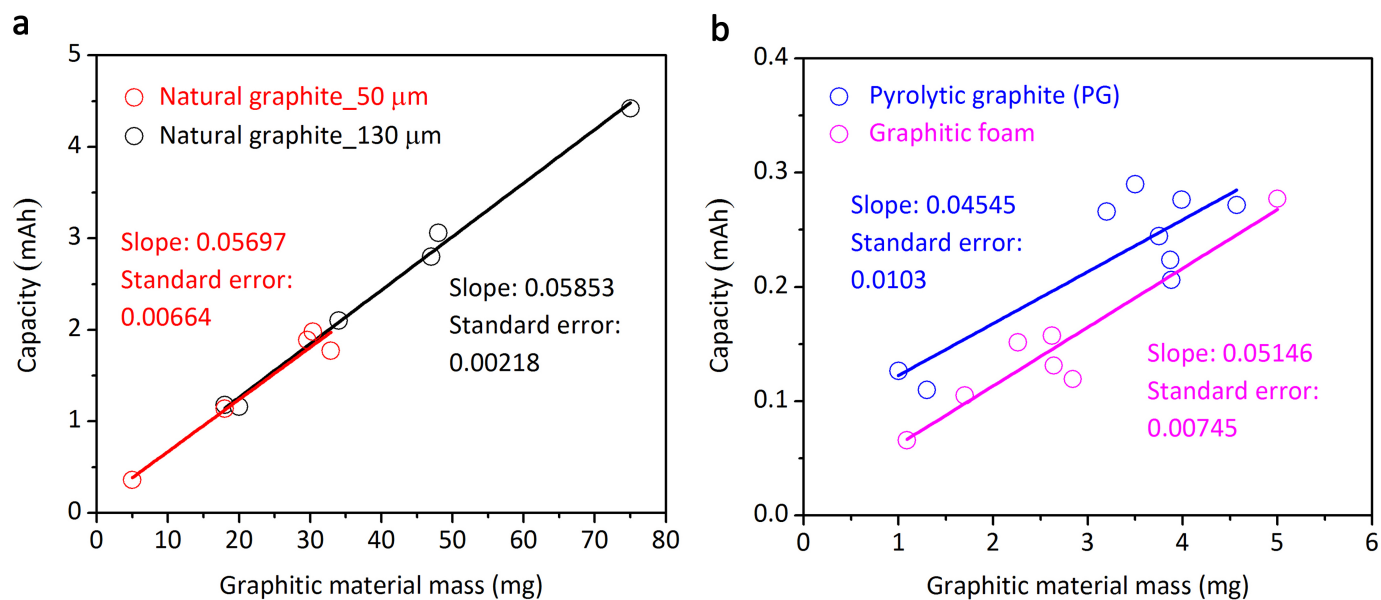


Extended Data Figure 1 | X-ray diffraction patterns of graphitic cathode materials. The natural graphite, pyrolytic graphite (PG) and graphitic foam exhibited typical graphite structure, with a sharp (002) X-ray diffraction (XRD) graphite peak at $2\theta \approx 26.55^\circ$ (d spacing = 3.35 Å).



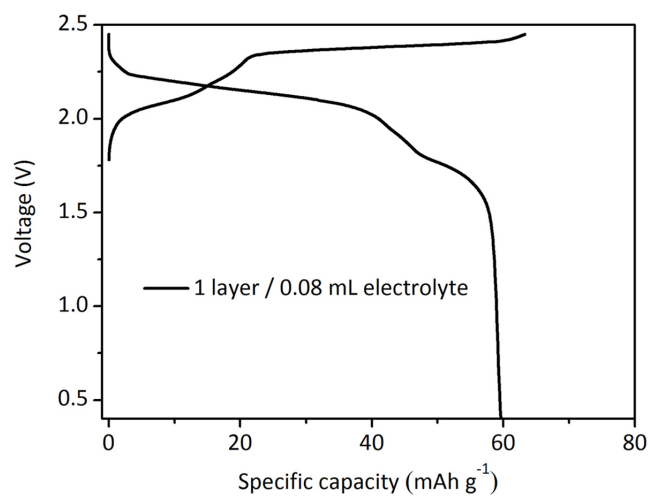
Extended Data Figure 2 | Determination of the optimal mole ratio of $\text{AlCl}_3/[\text{EMIm}]\text{Cl}$ ionic liquid electrolyte. **a**, Galvanostatic charge and discharge curves of Al/PG cells at a current density of 66 mA g^{-1} in various mole ratios (1.1, 1.3, 1.5 and 1.8) of $\text{AlCl}_3/[\text{EMIm}]\text{Cl}$ ionic liquid electrolytes

in a Swagelok-type cell. The Coulombic efficiencies of the cells are shown in parentheses. **b**, Raman spectrum of the ionic liquid electrolyte with a mole ratio of $\text{AlCl}_3/[\text{EMIm}]\text{Cl} = 1.3$.

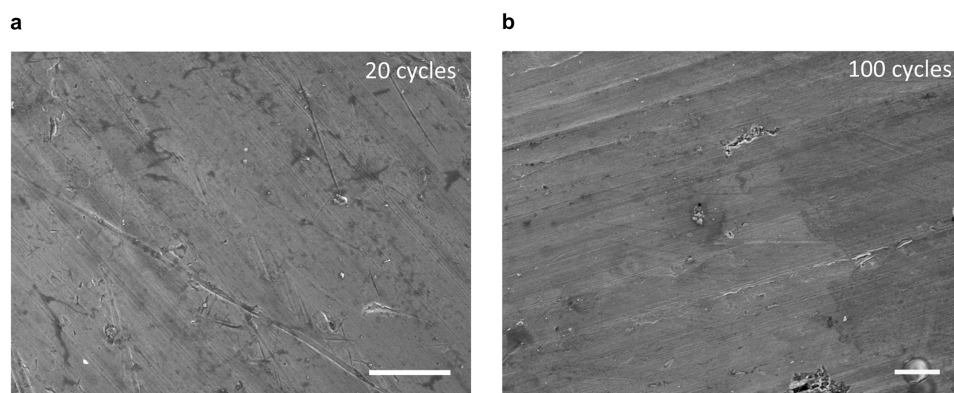


Extended Data Figure 3 | Calculated discharging capacities of Al/graphite cells with different masses of graphitic materials. a, Natural graphite foils of 50 μm and 130 μm thickness; b, PG and graphitic foam. These data suggest

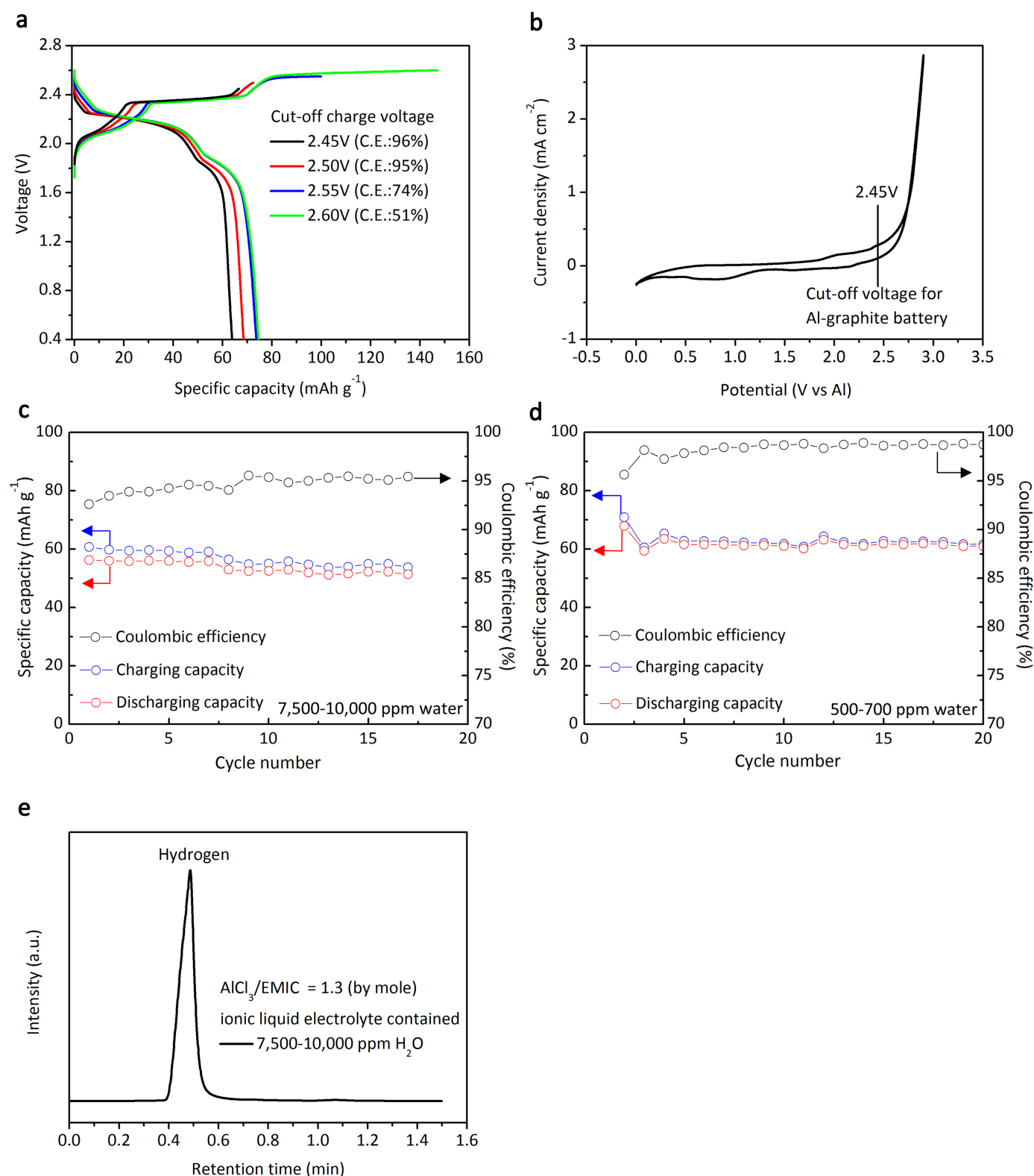
that the entire graphitic material (natural graphite, PG and graphitic foam) participated in the cell cathode reaction.



Extended Data Figure 4 | Galvanostatic charge and discharge curves of an Al/PG cell. The cell was constructed with one layer of glass fibre separator and 0.08 ml of ionic liquid electrolyte, suggesting that the minimum amount of electrolyte could be 0.02 ml per mg of PG. This electrochemical study was performed in an ionic liquid electrolyte of composition $\text{AlCl}_3/[\text{EMIm}]\text{Cl} = 1.3$ (by mole) at a current density of 66 mA g^{-1} in a Swagelok-type cell.

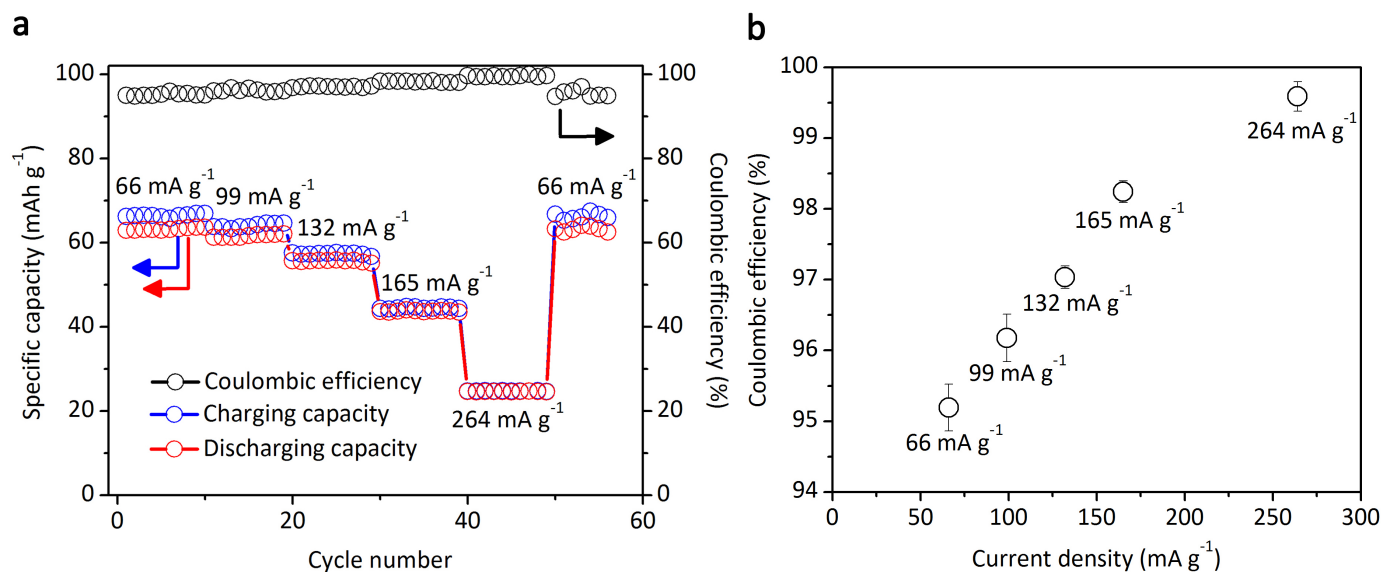


Extended Data Figure 5 | Surface observations of an Al anode. **a, b,** SEM images of the Al anode obtained from two Al/Pg cells after 20 (**a**) and 100 (**b**) cycles, respectively, and indicate no dendrite formation over these cycles. Scale bars, 10 μm.



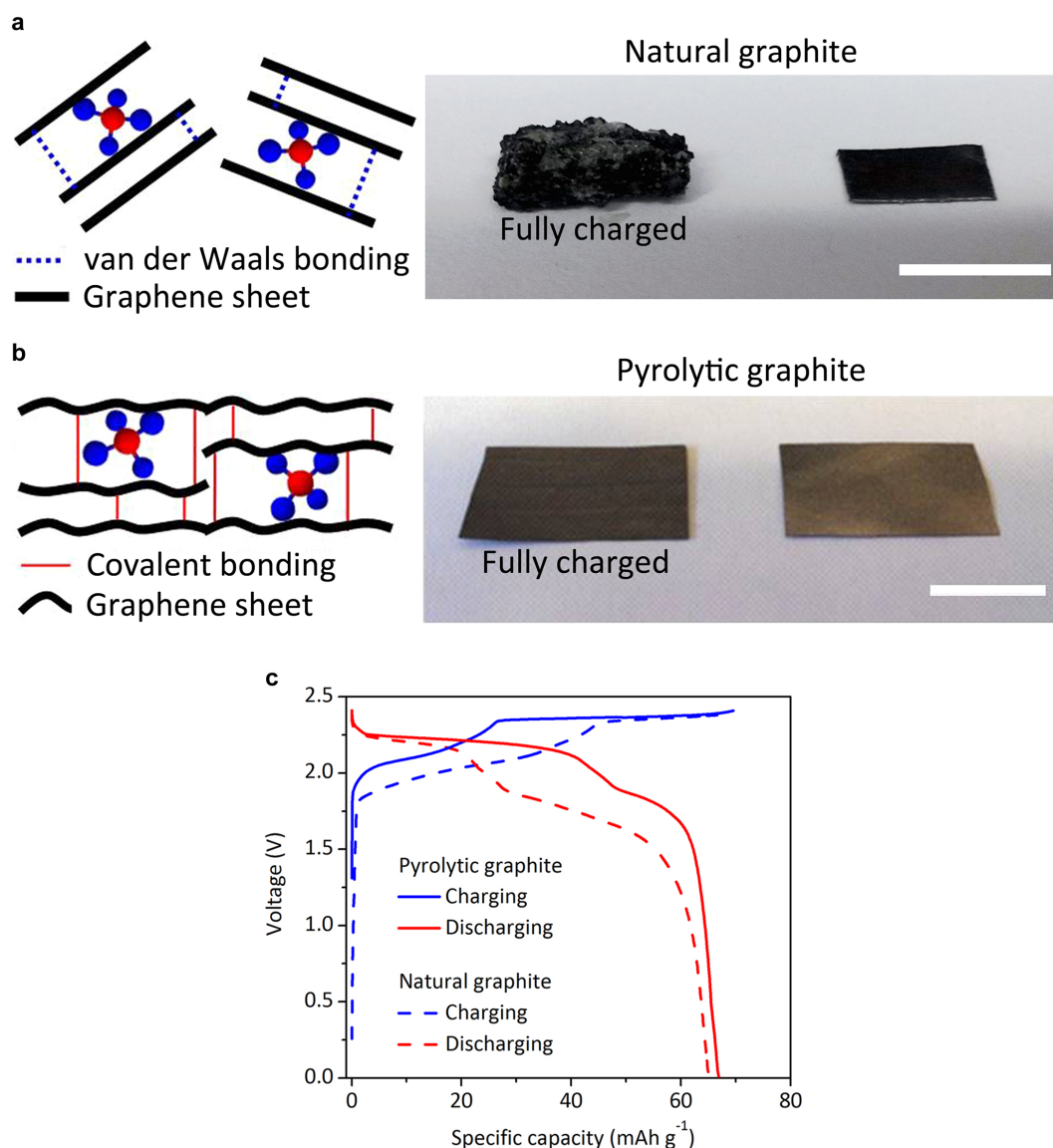
Extended Data Figure 6 | Electrochemical stability of the $\text{AlCl}_3/[\text{EMIm}]\text{Cl}$ ionic liquid electrolyte. **a**, Galvanostatic curves of Al/PG cells with different cut-off charge voltages obtained at 66 mA g^{-1} in a Swagelok-type cell. **b**, Cyclic voltammetry curve of a Al/glassy carbon (GC) cell at 10 mV s^{-1} in a Swagelok-type cell. **c**, **d**, Stability test of Al/natural graphite pouch cell at 66 mA g^{-1} in electrolytes containing water at 7,500–10,000 p.p.m. (**c**) and 500–700 p.p.m. (**d**). The Coulombic efficiencies are respectively 95.2% and

98.6%, and the discharge capacities are respectively 54.9 and 61.8 mA h g^{-1} at the 15 cycle. **e**, Gas chromatography spectrum of gaseous samples withdrawn from Al/graphite cells after 30 cycles using electrolyte with 7,500–10,000 p.p.m. H_2O content. The peak found in the retention time at $\sim 0.5 \text{ min}$ corresponds to hydrogen gas and matches the retention time of pure hydrogen gas used for calibration.



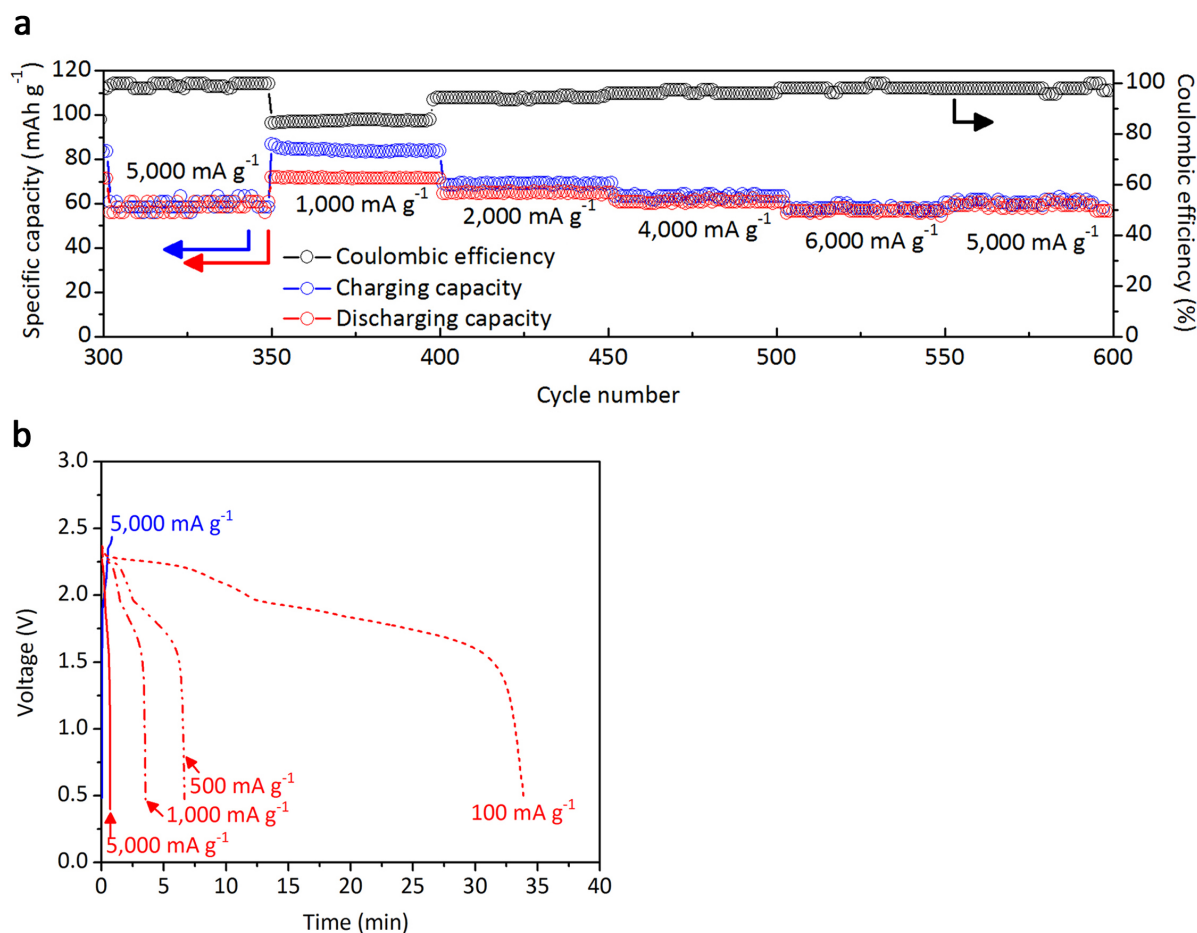
Extended Data Figure 7 | Rate capability of an Al/PG cell. **a**, Capacity retention of an Al/PG cell cycled at various current densities, showing good cycling stability at different charge–discharge current densities. **b**, Coulombic efficiency versus current density data of Al/PG cells, indicating the Coulombic

efficiency is ~ 95 – 97% at current densities of 66 – 132 mA g^{-1} . Error bars, standard deviation from the Coulombic efficiency for each current density. All electrochemical studies were performed in an ionic liquid electrolyte of composition $\text{AlCl}_3/[\text{EMIm}]\text{Cl} = 1.3$ (by mole) in a Swagelok-type cell.



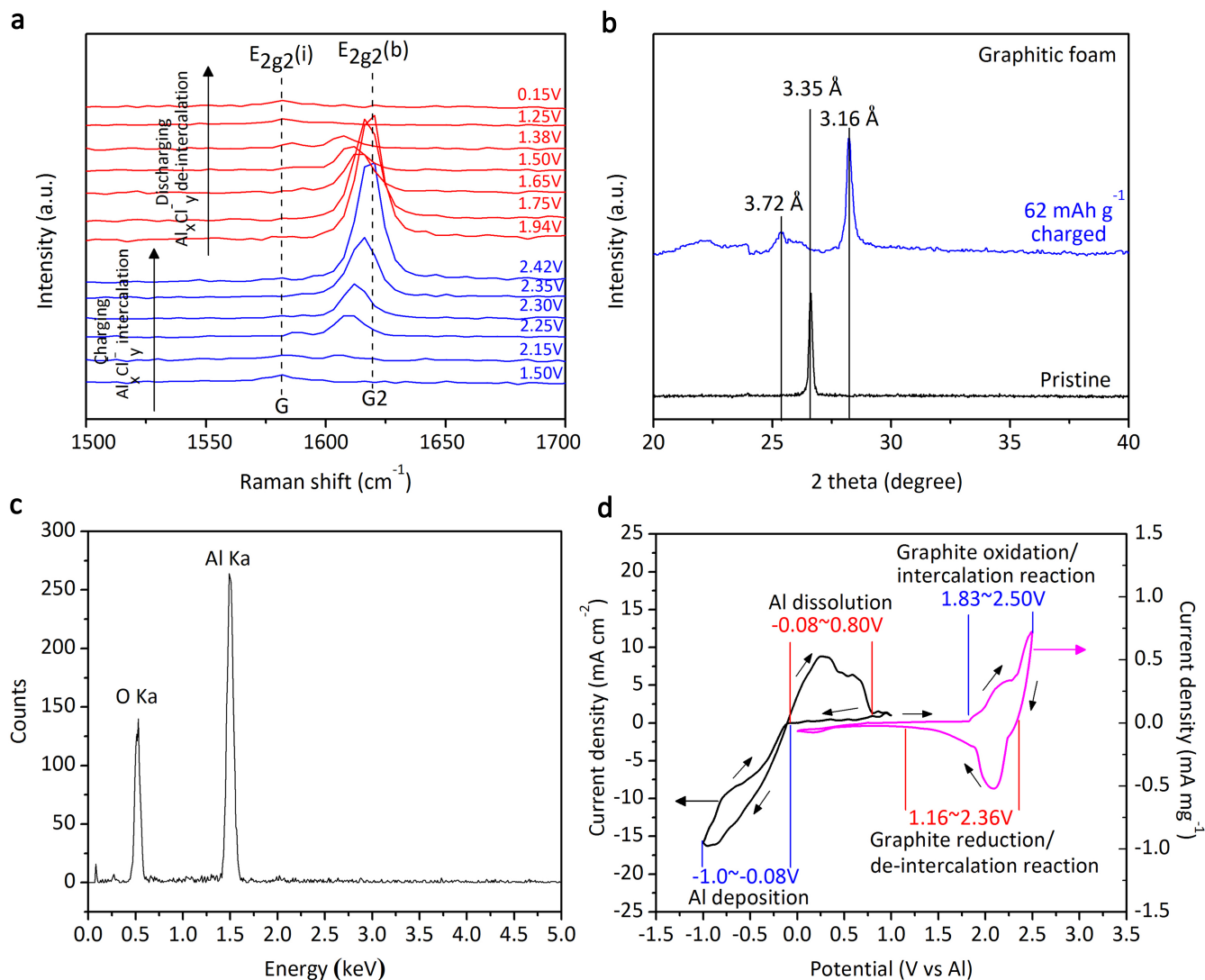
Extended Data Figure 8 | Advantages of PG as the cathode for an Al/graphite cell. **a, b,** Right: photographs of natural graphite (**a**) and pyrolytic graphite (PG; **b**) before and after being fully charged in an $\text{AlCl}_3/[\text{EMIm}]\text{Cl} = 1.3$ (by mole) ionic liquid electrolyte. Scale bars, 1 cm. Left: the schematic plots indicate the chemical bonds between the graphene

sheets of natural graphite (Vander Waals bonding) and of PG (covalent bonding). **c,** Galvanostatic charge and discharge curves of an Al/PG cell (at 66 mA g^{-1}) and an Al/natural graphite cell (at 33 mA g^{-1}) in an ionic liquid electrolyte of composition $\text{AlCl}_3/[\text{EMIm}]\text{Cl} = 1.3$ (by mole) in a Swagelok-type cell.



Extended Data Figure 9 | Rate capability of an Al/graphitic-foam cell.
a, Capacity retention of an Al/graphitic-foam cell cycled at various current densities, showing cycling stability at different charge–discharge current densities. All electrochemical studies were performed in an

$\text{AlCl}_3/[\text{EMIm}]\text{Cl} = 1.3$ (by mole) ionic liquid electrolyte in a pouch cell.
b, Galvanostatic charge and discharge curves of Al/graphitic-foam cells charging at $5,000 \text{ mA g}^{-1}$ and discharging at various current densities ranging from 100 to $5,000 \text{ mA g}^{-1}$. Same electrolyte and cell type as **a**.



Extended Data Figure 10 | Reaction mechanism of graphitic cathodes.

a, In situ Raman spectra recorded for the graphitic-foam cathode through a charge/discharge cycle showing chloroaluminate anion intercalation/de-intercalation into graphite. **b**, Ex situ XRD patterns of the pristine and fully

charged (62 mA h g⁻¹) graphitic foam. **c**, EDS spectrum of as-calcined fully charged (62 mA h g⁻¹) PG at 850 °C in air. **d**, Cyclic voltammetry curves of Al foil and PG at a scan rate of 10 mV s⁻¹ against an Al reference electrode.

A pomegranate-inspired nanoscale design for large-volume-change lithium battery anodes

Nian Liu^{1†}, Zhenda Lu^{2†}, Jie Zhao², Matthew T. McDowell², Hyun-Wook Lee², Wenting Zhao² and Yi Cui^{2,3*}

Silicon is an attractive material for anodes in energy storage devices^{1–3}, because it has ten times the theoretical capacity of its state-of-the-art carbonaceous counterpart. Silicon anodes can be used both in traditional lithium-ion batteries and in more recent Li–O₂ and Li–S batteries as a replacement for the dendrite-forming lithium metal anodes. The main challenges associated with silicon anodes are structural degradation and instability of the solid-electrolyte interphase caused by the large volume change (~300%) during cycling, the occurrence of side reactions with the electrolyte, and the low volumetric capacity when the material size is reduced to a nanometre scale^{4–7}. Here, we propose a hierarchical structured silicon anode that tackles all three of these problems. Our design is inspired by the structure of a pomegranate, where single silicon nanoparticles are encapsulated by a conductive carbon layer that leaves enough room for expansion and contraction following lithiation and delithiation. An ensemble of these hybrid nanoparticles is then encapsulated by a thicker carbon layer in micrometre-size pouches to act as an electrolyte barrier. As a result of this hierarchical arrangement, the solid-electrolyte interphase remains stable and spatially confined, resulting in superior cyclability (97% capacity retention after 1,000 cycles). In addition, the microstructures lower the electrode–electrolyte contact area, resulting in high Coulombic efficiency (99.87%) and volumetric capacity (1,270 mAh cm^{–3}), and the cycling remains stable even when the areal capacity is increased to the level of commercial lithium-ion batteries (3.7 mAh cm^{–2}).

Particle fracture and loss of electrical contact have long been identified as primary reasons for the capacity fading of silicon-based anodes. Pioneering works have shown that decreasing the feature size to the nanoscale allows for the material to withstand the large (de)lithiation strains without fracture^{8–16}. However, the cycle life of nanosized silicon is still limited due to the unstable solid-electrolyte interphase (SEI) on the surface. At the working potential of anodes (<0.5 V versus Li/Li⁺), the organic electrolyte decomposes and forms a thin SEI layer^{17–19}. When the silicon expands and contracts, the SEI layer deforms and breaks. The formation of new SEI on the freshly exposed silicon surface causes the cell to have poor Coulombic efficiency, with the accumulated SEI eventually blocking Li⁺ transport. As a consequence of this mechanism, even if most of the silicon active material remains electrically connected, the capacity decays as the SEI thickens. To control SEI formation, an electrolyte blocking layer and an internal void space need to be engineered into the structure. Such a design has been demonstrated in a few examples, including double-walled silicon nanotube and Si–C yolk–shell structures^{20–25}, yielding

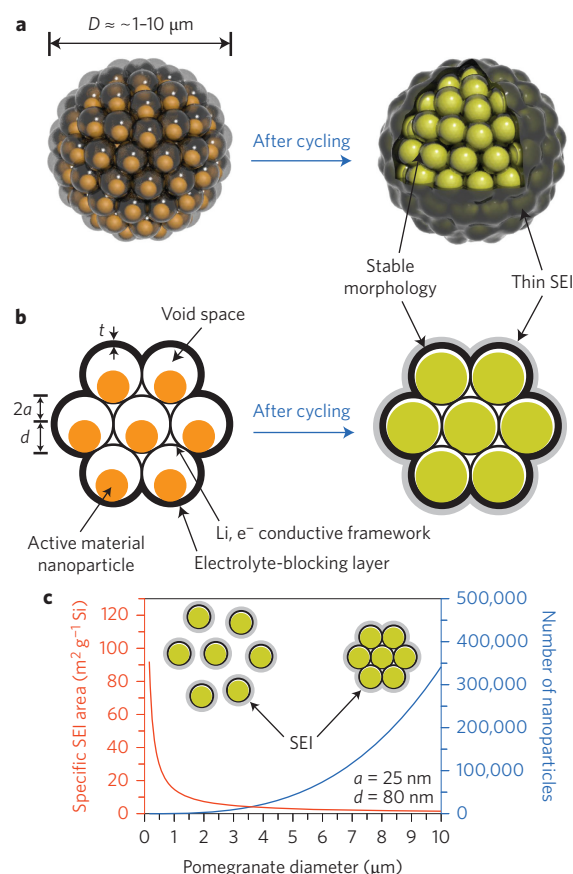


Figure 1 | Schematic of the pomegranate-inspired design. a,b, Three-dimensional view (**a**) and simplified two-dimensional cross-section view (**b**) of one pomegranate microparticle before and after electrochemical cycling (in the lithiated state). The nanoscale size of the active-material primary particles prevents fracture on (de)lithiation, whereas the micrometre size of the secondary particles increases the tap density and decreases the surface area in contact with the electrolyte. The self-supporting conductive carbon framework blocks the electrolyte and prevents SEI formation inside the secondary particle, while facilitating lithium transport throughout the whole particle. The well-defined void space around each primary particle allows it to expand without deforming the overall morphology, so the SEI outside the secondary particle is not ruptured during cycling and remains thin. **c**, Calculated surface area in contact with electrolyte (specific SEI area) and the number of primary nanoparticles in one pomegranate particle versus its diameter.

¹Department of Chemistry, Stanford University, Stanford, California 94305, USA, ²Department of Materials Science and Engineering, Stanford University, Stanford, California 94305, USA, ³Stanford Institute for Materials and Energy Sciences, SLAC National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, California 94025, USA, [†]These authors contributed equally to this work. *e-mail: yicui@stanford.edu

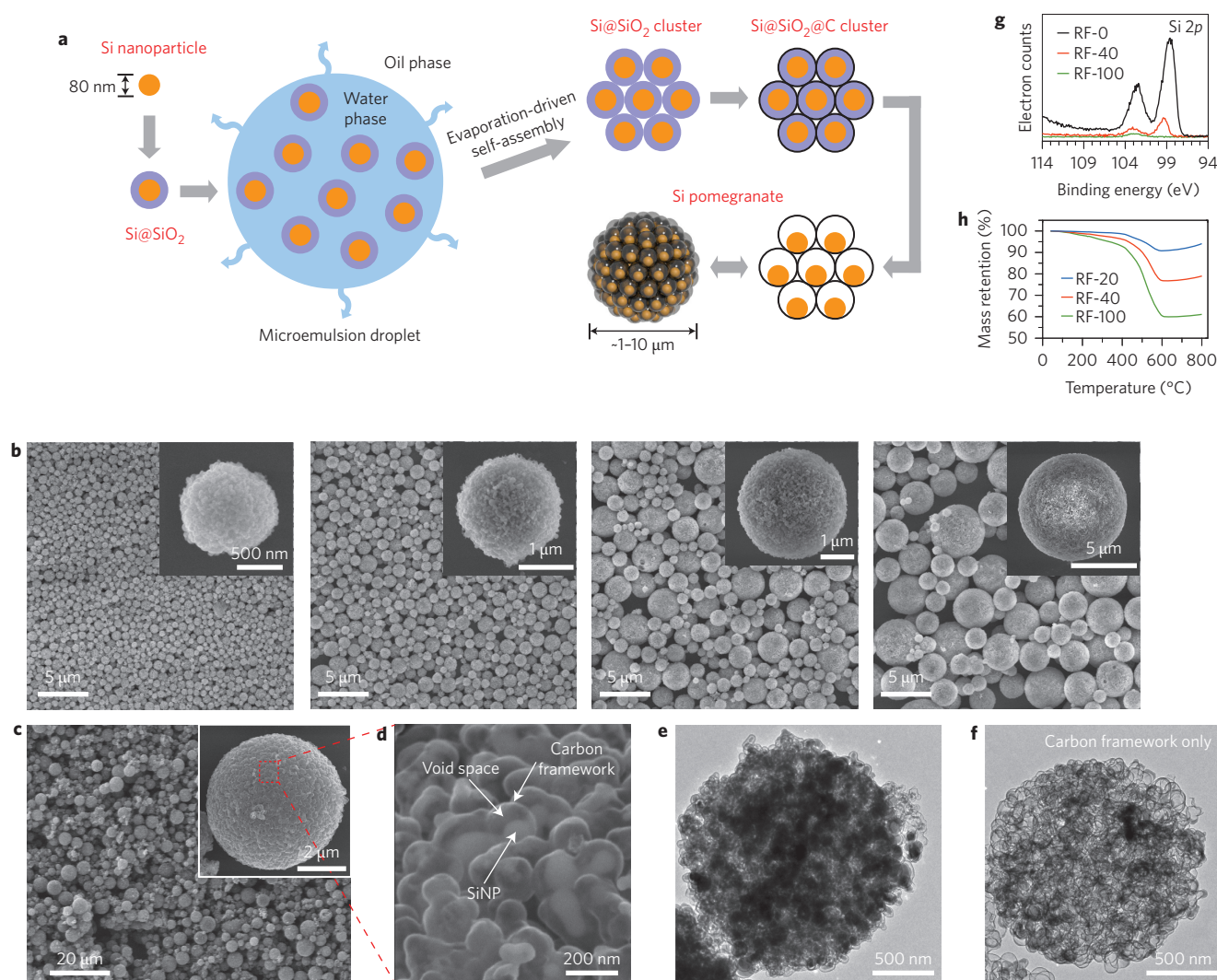


Figure 2 | Fabrication and characterization of silicon pomegranates. **a**, Schematic of the fabrication process for silicon pomegranates. **b**, SEM images of a series of silicon nanoparticle clusters with different diameters assembled by the microemulsion approach. The insets show magnified images of the individual microparticles. **c**, SEM images of silicon pomegranates showing the micrometre-sized and spherical morphology. **d**, Magnified SEM image showing the local structure of silicon nanoparticles and the conductive carbon framework with well-defined void space between. **e**, TEM image of one silicon pomegranate particle. **f**, TEM image of the carbon framework after etching away silicon using NaOH. **g,h**, High-resolution XPS spectra of Si 2p peaks (**g**) and thermogravimetric profiles (**h**) of silicon pomegranates with different thicknesses of carbon framework. X in 'RF-X' denotes X mg of resorcinol used for each 100 mg of Si@SiO₂ in the RF coating step (see Supplementary Methods). RF-0 denotes silicon nanoparticle clusters without carbon.

significant improvement in cycle life with relatively low areal mass loading ($<0.2 \text{ mg cm}^{-2}$).

Although nanostructuring has been successful in extending the cycle life of silicon, nanostructured electrodes have introduced new fundamental challenges, including higher surface area, low tap density and generally poor electrical properties due to the higher interparticle resistance. The high surface area increases side reactions with the electrolyte and lowers the Coulombic efficiency. The low tap density leads to low volumetric capacity and a thick electrode at high mass loading, which makes it difficult to maintain electrical and ionic pathways during cycling. Finally, electrical contact between the nanoparticles is easily altered or diminished by volume changes during cycling, severely decreasing the cycle life of the electrode. To the best of our knowledge, stable cycling (100 cycles) with an areal capacity higher than 3 mAh cm^{-2} has not been reported.

Here, inspired by the structure of a pomegranate fruit, we demonstrate a novel secondary structure for silicon anodes (Fig. 1a,b). Such a design has multiple advantages: (1) the nanosized

primary particle size prevents fracture; (2) the well-defined internal void space allows the silicon to expand without changing the secondary particle size; (3) the carbon framework functions as an electrical highway and a mechanical backbone so that all nanoparticles are electrochemically active; (4) carbon completely encapsulates the entire secondary particle, limiting most SEI formation to the outer surface instead of on individual nanoparticles, which not only limits the amount of SEI, but also retains the internal void space for silicon expansion; and (5) the dilemma of high surface area and low tap density introduced when using nanosized primary features is partially solved. With this design, the specific SEI area (the surface area in contact with the electrolyte divided by the mass of silicon) decreases from $\sim 90 \text{ m}^2 \text{ g}^{-1}$ for single yolk-shell particles to $\sim 15 \text{ m}^2 \text{ g}^{-1}$ for $1 \mu\text{m}$ secondary particles and to only $1.5 \text{ m}^2 \text{ g}^{-1}$ for $10 \mu\text{m}$ particles (Fig. 1c). At the same time, because of the space-efficient packing inside the secondary particles, their tap density is significantly higher than primary nanosized particles packed randomly. In the simplified demonstration shown in Supplementary Fig. 1, silicon nanoparticles have a tap density of

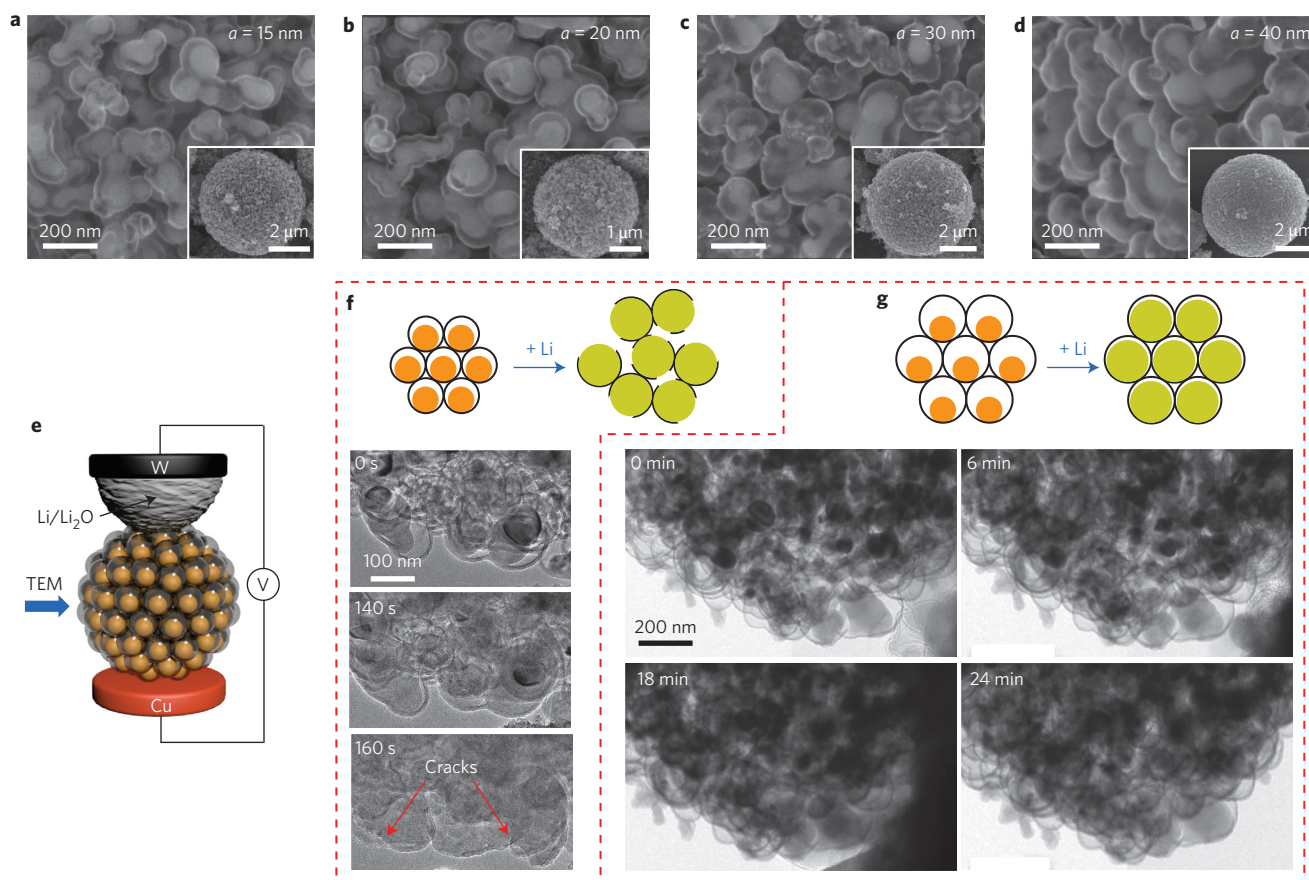


Figure 3 | Tuning the size of the void space of silicon pomegranates and *in situ* TEM characterization during lithiation. **a–d**, SEM images of as-synthesized silicon pomegranates with a void space size of 15 nm (**a**), 20 nm (**b**), 30 nm (**c**) and 40 nm (**d**), respectively. The definition of void space size is shown in Fig. 1b. **e**, Schematic of the *in situ* TEM device. **f**, Schematic and time-lapse images of the lithiation of silicon pomegranates with insufficient (~15 nm) void space (Supplementary Movie 1). Lithium transports along and across the carbon framework to react with the silicon inside, causing volume expansion. Because the void space is insufficient, the carbon framework is ruptured by the expansion of the silicon, and the overall morphology is destroyed. **g**, Lithiation of a silicon pomegranate with sufficient (~40 nm) void space (Supplementary Movie 2); the nanoparticles expand within the carbon framework and the carbon framework does not rupture. The secondary particle morphology is therefore intact on lithiation.

0.15 g cm⁻², while micrometre-sized clusters have a tap density of 0.53 g cm⁻², which represents an increase of 250%. This tap density is within the desirable range for electrodes because the volume expansion of silicon is up to 300%.

We have developed a bottom-up microemulsion²⁶ approach to synthesize highly spherical silicon pomegranate microbeads ranging from 500 nm to 10 μm in diameter (Fig. 2a,b, Supplementary Figs 2–5). Silicon nanoparticles are individually encapsulated by the carbon framework, with well-defined void spaces between the silicon and carbon (Fig. 2c–e) that accommodate the volume expansion of the silicon. The carbon framework, although only a few nanometres thick, firmly supports the whole microbead (Fig. 2f, Supplementary Figs 6 and 7). X-ray photoelectron spectroscopy (XPS) results indicate complete coverage of carbon (Fig. 2g), which is crucial for blocking the electrolyte and limiting the majority of SEI formation to the outer surface of the microbeads. Our unique pomegranate design is fundamentally different from previous reports of silicon/carbon composite secondary particles^{11,27,28}. First, the carbon component functions not only as a conducting framework, but also as an electrolyte-blocking layer, so SEI forms mostly outside the secondary particle. Second, the void spaces inside the secondary particle are well-defined and evenly distributed around each nanoparticle, and effectively accommodate the volume expansion of the silicon without rupturing the carbon shell or changing the secondary particle size. The initial well-defined void spaces do not decrease

the volumetric capacity, because they are designed to be mostly occupied in the lithiated state.

A sufficient and well-defined internal void space is necessary to maintain the structural integrity of the silicon anode. To identify the critical gap size, we synthesized silicon pomegranates with a series of gap sizes by tuning the thickness of the SiO₂ sacrificial layer (Fig. 3a–d). We then performed an *in situ* transmission electron microscopy (TEM) study of their structural change on lithiation. The *in situ* TEM set-up is based on previous studies^{29,30} and is shown schematically in Fig. 3e. For a silicon pomegranate with insufficient (~15 nm) gap size, as shown in Fig. 3f and Supplementary Movie 1, the expanded Li–Si first occupies all the void spaces and then ruptures the carbon shell and causes structural degradation. This kind of crack formation in a conventional battery configuration will cause (1) excessive formation of SEI, (2) loss of electrical connection to the active material and (3) whole-electrode-level cracking due to accumulated particle shuffling across the thickness of the electrode. In the case of sufficient gap size (Fig. 3g and Supplementary Movie 2), however, silicon expands inside the carbon framework to occupy the void spaces, resulting in little change to both the carbon shell and secondary particle size. Taking into account the initial size of the nanoparticles (~80 nm) and their size distribution, silicon pomegranates with a ~30–40 nm gap size are the most promising, so this gap size was chosen for further electrochemical characterization. The fact that the nanoparticles in the middle of the secondary particle are

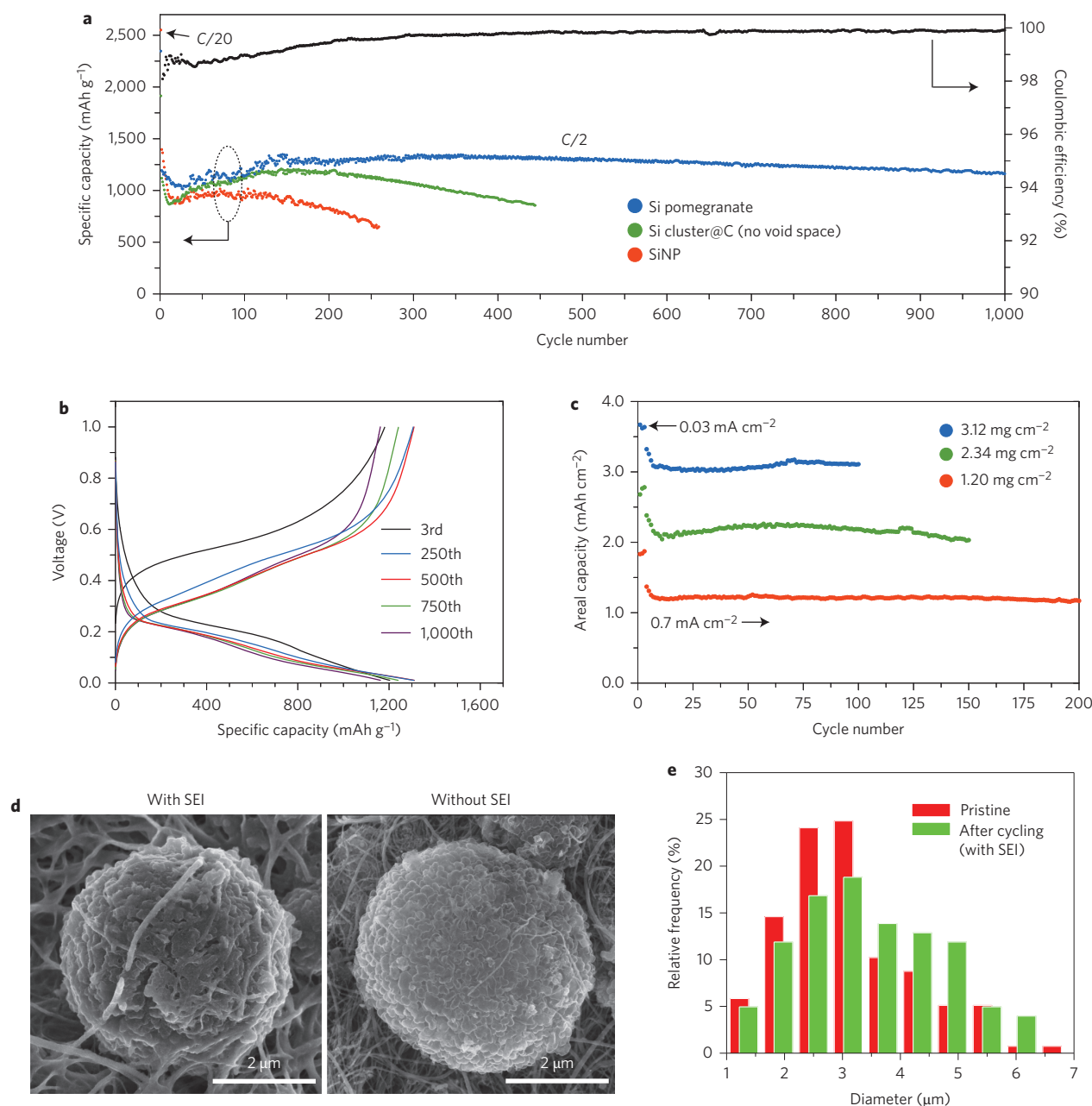


Figure 4 | Electrochemical characterization of silicon pomegranate anodes. All specific capacities of silicon pomegranate anodes are based on the total mass of the active materials (silicon and carbon in the pomegranate structure). **a**, Reversible delithiation capacity for the first 1,000 galvanostatic cycles of the silicon pomegranate and other structures tested under the same conditions. Coulombic efficiency is plotted for the silicon pomegranate only. The mass loading of the active material was $\sim 0.2 \text{ mg cm}^{-2}$. The rate was C/20 for the first cycle and C/2 for later cycles. ($1C = 4.2 \text{ A g}^{-1}$ active material). **b**, Voltage profiles for the silicon pomegranate plotted for the 3rd, 250th, 500th, 750th and 1,000th cycles. **c**, High areal mass loading test (up to 3.12 mg cm^{-2} active material) of silicon pomegranate anodes. All electrodes were first cycled at 0.03 mA cm^{-2} for three cycles and 0.7 mA cm^{-2} for later cycles. **d**, Typical SEM images of silicon pomegranates after 100 cycles. **e**, Statistical analysis of silicon pomegranate diameter before and after 100 cycles, with averages of 3.1 and 3.4 μm , respectively.

active upon lithiation suggests effective lithium transport along and across the carbon shell and framework.

The pomegranate design affords remarkable battery performance. As shown in Fig. 4a, its reversible capacity reached $2,350 \text{ mAh g}^{-1}$ at a rate of C/20 ($1C = \text{charge/discharge in 1 h}$). If not mentioned, all reported capacities are based on the total mass of silicon and carbon in the pomegranate structure. Because silicon is only 77% of the mass of the pomegranate structure, the capacity with respect to silicon is as high as $3,050 \text{ mAh g}^{-1}$. The volumetric capacity based on electrode volume was determined

to be $1,270 \text{ mAh cm}^{-3}$, which is more than twice the 600 mAh cm^{-3} obtained for graphite anodes⁶. From the 2nd to 1,000th cycle at a rate of C/2, the capacity retention was more than 97%. After 1,000 cycles, over $1,160 \text{ mAh g}^{-1}$ capacity remained, which is more than three times the theoretical capacity of graphite. The cycle stability (0.003% decay per cycle) is among the best cycling performances of silicon anodes reported to date. Furthermore, it was achieved with a conventional polyvinylidene fluoride (PVDF) binder, which has been considered a poor binder for silicon anodes¹³. Under the same conditions, secondary particles

without an internal void space (nanoparticle clusters directly coated by carbon) demonstrated significant decay after 200 cycles. Bare nanoparticles decayed even more quickly. The voltage profiles of silicon pomegranate electrodes (Fig. 4b) exhibited typical electrochemical features of silicon, with little change over 1,000 cycles. Coulombic efficiency is an indicator of the reversibility of the electrode reaction. SEI rupture and reformation usually results in decreased Coulombic efficiency, especially in later cycles. The average Coulombic efficiency from the 500th to 1,000th cycles of the silicon pomegranate is as high as 99.87% (Fig. 4a, Supplementary Fig. 8). At a relatively slow rate of $C/2$, this Coulombic efficiency is superior to most previous reports.

It should be noted that many publications only report specific capacity normalized by the weight of the active materials, and low areal mass loading is frequently used to achieve stable cycling. However, high areal mass loading is needed to realize high performance based on total cell weight or volume³¹. We therefore tested thick silicon pomegranate electrodes with active material mass loading up to 3.12 mg cm^{-2} (silicon and carbon in a pomegranate structure). Following deep cycling at 0.03 mA cm^{-2} , the reversible areal capacity reached 3.67 mAh cm^{-2} (Fig. 4c), similar to, if not higher than, the capacity in a commercial lithium-ion battery cell. From the 4th to 100th cycle at a higher rate of 0.7 mA cm^{-2} , the capacity retention was as high as 94%. After 100 deep cycles, the areal capacity was still above 3 mAh cm^{-2} . To the best of our knowledge, high capacity and stable cycling at such a mass loading level have rarely been reported for silicon anodes. At slightly lower mass loadings of 2.34 and 1.20 mg cm^{-2} , 150 and 200 stable cycles were achieved. The specific capacity of the high mass loading cell (950 mAh g^{-1}) is only slightly lower than that of the low mass loading cell shown in Fig. 4a, which indicates that almost all the silicon pomegranate is active in the thick electrode. Further high loading cells also demonstrated stable cycling (Supplementary Fig. 9). Stable cycling of a high mass loading cell has strict requirements in terms of structural stability at the particle level, because even small changes in particle morphology could accumulate across the thickness of the electrode and cause electrode-level cracking and failure. Excellent performance at high mass loading indicates successful design of the pomegranate structure.

There are two interdependent characteristics of the pomegranate design that enable superior battery performance. The first is the internally accommodated volume expansion, which retains the structural integrity of the secondary particles and stabilizes the SEI on the surface. After 100 deep cycles, the morphology of the silicon pomegranates was examined with scanning electron microscopy (SEM; Fig. 4d). Nearly perfect spherical micrometre-sized particles were observed, with thin and uniform SEI coating the surface. After removing the SEI with acid, intact carbon shells with silicon inside are clearly visible. Statistical analysis of the diameter of the silicon pomegranates before and after cycling (Fig. 4e) shows only a 10% increase from $3.1 \mu\text{m}$ to $3.4 \mu\text{m}$ due to the formation of an SEI layer. The volume change here is much less than that of bare alloying anode, and is as small as intercalation-type graphite particles⁶, which is crucial for excellent performance in volume change-sensitive, high mass loading cells.

The second characteristic is spatially confined SEI formation. Our pomegranate particles are fundamentally different from previously reported secondary particles that have open inner surfaces^{11,27,28}, in that they have an electrolyte-blocking layer that limits most of the SEI formation to the surface of the secondary particle (Supplementary Figs 10 and 11). Also, the internally accommodated volume expansion ensures that the SEI on the secondary particle is thin and stable. This mechanism not only decreases the quantity of SEI, but also enables the inner void space to be retained,

even after many cycles. As shown in Fig. 4e, the thickness of the outer SEI layer is only $\sim 150 \text{ nm}$, similar to that of double-walled silicon nanotubes²⁰. In addition, because the SEI is on a micrometre-sized secondary particle, the relative quantity of SEI is significantly reduced compared to that on primary nanostructures (Fig. 1c). The thin, stable and spatially confined SEI increases the Coulombic efficiency and improves cycle stability.

An additional advantage of silicon pomegranates is that they exist in the form of powders, and their synthesis does not involve any complex equipment or processes such as chemical vapour deposition. Hence, they are entirely compatible with conventional slurry-coating manufacturing for lithium-ion battery electrodes. However, we note that future work is needed to reduce the cost of starting materials such as the silicon nanoparticles to meet commercial needs. The initial Coulombic efficiency is 82% for 9% carbon, but decreases with increasing thickness of the carbon coating (75% Coulombic efficiency at 23% carbon) (Fig. 2h, Supplementary Fig. 12), and is still lower than state-of-the-art graphite anodes due to the high density of lithium trapping sites in amorphous carbon (Supplementary Fig. 13). It still needs to be improved further, possibly by performing prelithiation³² or replacing amorphous carbon with another material that does not irreversibly trap large amounts of lithium.

Methods

Synthesis. As illustrated in Fig. 2a, commercial silicon nanoparticles were first coated with a SiO_2 layer using tetraethoxysilane²¹. The aqueous dispersion of Si@SiO_2 nanoparticles was then mixed with 1-octadecene containing 0.3 wt% emulsifier²⁶ to form water-in-oil emulsions. After evaporation of water at $\sim 95\text{--}98^\circ\text{C}$, the assembled Si@SiO_2 nanoparticle clusters were collected by centrifugation, followed by heat treatment at 550°C for 1 h in air to remove the organics and condense the cluster structures. A low-cost, step-growth polymerization in the presence of ammonia then generated a resorcinol-formaldehyde resin (RF) layer to wrap the cluster, which was converted into a carbon layer under argon at 800°C . The thickness of the carbon layer could be tuned by changing the added amount of resorcinol monomer³³. Finally, the SiO_2 sacrificial layer was removed with 5 wt% HF solution to form a void space to accommodate the large volume change of silicon material during the charge/discharge process. For more details, see Supplementary Materials and Methods.

In situ TEM. A specialized dual-probe electrical biasing holder (Nanofactory Instruments) was used. By biasing the working electrode between -2.5 and -3 V versus the counter electrode, Li^+ ions flow through the oxide/nitride layer and are reduced at the working electrode, where they react with carbon and alloy with the silicon in the pomegranate structure. The lithiation time of a silicon pomegranate structure is less than 30 min (Fig. 3g), corresponding to $2C$. The small size of the primary nanoparticles and the conductive carbon framework synergistically lead to good kinetics, as also evidenced by the rate performance of half cells shown in Supplementary Fig. 14.

Electrochemistry. The battery performance was evaluated by galvanostatic cycling of coin cells with the silicon pomegranate structures as the working electrode and lithium foil as the counter/reference electrode. The working electrodes were made using a typical slurry method with silicon pomegranate powders, carbon black and PVDF binder with a mass ratio of 8:1:1; the mass loading of active material (silicon and carbon in the pomegranate structure) was $\sim 0.2 \text{ mg cm}^{-2}$. All the cells were cycled between 0.01 and 1 V versus Li/Li^+ . To prepare high mass loading electrodes, as shown in Fig. 4c, silicon pomegranate microbeads and carbon nanotubes (mass ratio 7:3) were dispersed in *N*-methyl-2-pyrrolidone (NMP) and filtered to make a binder-free microbead/carbon nanotube paper, which was cut into discs, with mass loading from 1 to 3 mg cm^{-2} . The electrolyte was 1.0 M LiPF_6 in 1:1 wt/wt ethylene carbonate/diethyl carbonate, with 1 vol% vinylene carbonate added to improve the cycling stability. This type of filtered electrode has a density of $\sim 0.4 \text{ g cm}^{-3}$. The thickness is $\sim 120 \mu\text{m}$ when the silicon pomegranate loading is 3.12 mg cm^{-2} . The volumetric capacity based on electrode volume is 310 mAh cm^{-3} . This value is lower than graphite, but it should be noted that such filtered electrodes are not calendared and 30% carbon nanotubes increases the volume of the electrode. They are intentionally not calendared to show the excellent and robust intraparticle conductivity of the silicon pomegranates without the need for pressure. To examine whether the silicon pomegranate could perform well in denser electrodes, we also tested high-loading electrodes fabricated using an industry-viable slurry coating and calendaring process with much fewer carbon nanotubes (5%) or no nanotubes. These electrodes also demonstrated stable cycling at high areal capacity (Supplementary Fig. 9), as well as much higher volumetric capacity ($\sim 900\text{--}1,270 \text{ mAh cm}^{-3}$) than the practical value for graphite anodes ($<600 \text{ mAh cm}^{-3}$).

Received 19 August 2013; accepted 10 January 2014;
published online 16 February 2014

References

- Armand, M. & Tarascon, J.-M. Building better batteries. *Nature* **451**, 652–657 (2008).
- Arico, A. S., Bruce, P., Scrosati, B., Tarascon, J.-M. & van Schalkwijk, W. Nanostructured materials for advanced energy conversion and storage devices. *Nature Mater.* **4**, 366–377 (2005).
- Bruce, P. G., Freunberger, S. A., Hardwick, L. J. & Tarascon, J.-M. Li–O₂ and Li–S batteries with high energy storage. *Nature Mater.* **11**, 19–29 (2012).
- Beaulieu, L. Y., Eberman, K. W., Turner, R. L., Krause, L. J. & Dahn, J. R. Colossal reversible volume changes in lithium alloys. *Electrochem. Solid-State Lett.* **4**, A137–A140 (2001).
- Obrovac, M. N. & Christensen, L. Structural changes in silicon anodes during lithium insertion/extraction. *Electrochem. Solid-State Lett.* **7**, A93–A96 (2004).
- Obrovac, M. N., Christensen, L., Le, D. B. & Dahn, J. R. Alloy design for lithium-ion battery anodes. *J. Electrochem. Soc.* **154**, A849–A855 (2007).
- Larcher, D. *et al.* Recent findings and prospects in the field of pure metals as negative electrodes for Li-ion batteries. *J. Mater. Chem.* **17**, 3759–3772 (2007).
- Chan, C. K. *et al.* High-performance lithium battery anodes using silicon nanowires. *Nature Nanotech.* **3**, 31–35 (2008).
- Wu, H. & Cui, Y. Designing nanostructured Si anodes for high energy lithium ion batteries. *Nano Today* **7**, 414–429 (2012).
- Park, M.-H. *et al.* Silicon nanotube battery anodes. *Nano Lett.* **9**, 3844–3847 (2009).
- Magasinski, A. *et al.* High-performance lithium-ion anodes using a hierarchical bottom-up approach. *Nature Mater.* **9**, 353–358 (2010).
- Deshpande, R., Cheng, Y.-T. & Verbrugge, M. W. Modeling diffusion-induced stress in nanowire electrode structures. *J. Power Sources* **195**, 5081–5088 (2010).
- Liu, G. *et al.* Polymers with tailored electronic structure for high capacity lithium battery electrodes. *Adv. Mater.* **23**, 4679–4683 (2011).
- Lee, S. W., McDowell, M. T., Berla, L. A., Nix, W. D. & Cui, Y. Fracture of crystalline silicon nanopillars during electrochemical lithium insertion. *Proc. Natl Acad. Sci. USA* **109**, 4080–4085 (2012).
- Hwang, T. H., Lee, Y. M., Kong, B.-S., Seo, J.-S. & Choi, J. W. Electrospun core-shell fibers for robust silicon nanoparticle-based lithium ion battery anodes. *Nano Lett.* **12**, 802–807 (2012).
- Yi, R., Dai, F., Gordin, M. L., Chen, S. & Wang, D. Micro-sized Si–C composite with interconnected nanoscale building blocks as high-performance anodes for practical application in lithium-ion batteries. *Adv. Energy Mater.* **3**, 295–300 (2013).
- Aurbach, D. Review of selected electrode–solution interactions which determine the performance of Li and Li ion batteries. *J. Power Sources* **89**, 206–218 (2000).
- Chan, C. K., Ruffo, R., Hong, S. S. & Cui, Y. Surface chemistry and morphology of the solid electrolyte interphase on silicon nanowire lithium-ion battery anodes. *J. Power Sources* **189**, 1132–1140 (2009).
- Verma, P., Maire, P. & Novák, P. A review of the features and analyses of the solid electrolyte interphase in Li-ion batteries. *Electrochim. Acta* **55**, 6332–6341 (2010).
- Wu, H. *et al.* Stable cycling of double-walled silicon nanotube battery anodes through solid-electrolyte interphase control. *Nature Nanotech.* **7**, 310–315 (2012).
- Liu, N. *et al.* A yolk-shell design for stabilized and scalable Li-ion battery alloy anodes. *Nano Lett.* **12**, 3315–3321 (2012).
- Li, X. *et al.* Hollow core–shell structured porous Si–C nanocomposites for Li-ion battery anodes. *J. Mater. Chem.* **22**, 11014–11017 (2012).
- Chen, S. *et al.* Silicon core–hollow carbon shell nanocomposites with tunable buffer voids for high capacity anodes of lithium-ion batteries. *Phys. Chem. Chem. Phys.* **14**, 12741–12745 (2012).
- Park, Y. *et al.* Si-encapsulating hollow carbon electrodes via electrodeless etching for lithium-ion batteries. *Adv. Energy Mater.* **3**, 206–212 (2012).
- Wang, B. *et al.* Contact-engineered and void-involved silicon/carbon nanohybrids as lithium-ion-battery anodes. *Adv. Mater.* **25**, 3560–3565 (2013).
- Cho, Y.-S., Yi, G.-R., Kim, S.-H., Pine, D. J. & Yang, S.-M. Colloidal clusters of microspheres from water-in-oil emulsions. *Chem. Mater.* **17**, 5006–5013 (2005).
- Yin, Y.-X., Xin, S., Wan, L.-J., Li, C.-J. & Guo, Y.-G. Electro spray synthesis of silicon/carbon nanoporous microspheres as improved anode materials for lithium-ion batteries. *J. Phys. Chem. C* **115**, 14148–14154 (2011).
- Jung, D. S., Hwang, T. H., Park, S. B. & Choi, J. W. Spray drying method for large-scale and high-performance silicon negative electrodes in Li-ion batteries. *Nano Lett.* **13**, 2092–2097 (2013).
- Huang, J. Y. *et al.* *In situ* observation of the electrochemical lithiation of a single SnO₂ nanowire electrode. *Science* **330**, 1515–1520 (2010).
- McDowell, M. T. *et al.* *In situ* TEM of two-phase lithiation of amorphous silicon nanospheres. *Nano Lett.* **13**, 758–764 (2013).
- De Volder, M. F. L., Tawfick, S. H., Baughman, R. H. & Hart, A. J. Carbon nanotubes: present and future commercial applications. *Science* **339**, 535–539 (2013).
- Liu, N., Hu, L., McDowell, M. T., Jackson, A. & Cui, Y. Prelithiated silicon nanowires as an anode for lithium ion batteries. *ACS Nano* **5**, 6487–6493 (2011).
- Li, N. *et al.* Sol-gel coating of inorganic nanostructures with resorcinol-formaldehyde resin. *Chem. Commun.* **49**, 5135–5137 (2013).

Acknowledgements

Y.C. acknowledges support from the Assistant Secretary for Energy Efficiency and Renewable Energy, Office of Vehicle Technologies of the US Department of Energy (contract no. DE-AC02-05CH11231, subcontract no. 6951379) under the Batteries for Advanced Transportation Technologies (BATT) Program. M.T.M. acknowledges the National Science Foundation Graduate Fellowship Program and the Stanford Graduate Fellowship Program. H.W.L. acknowledges the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (contract no. 2012038593). The authors thank Z. Chen for discussions, F. Wei for providing the carbon nanotubes and Croda for providing the emulsifier.

Author contributions

N.L., Z.L. and Y.C. conceived the concept and experiments. N.L. and Z.L. carried out the synthesis and performed materials characterization and electrochemical measurements. J.Z. participated in part of the synthesis and electrochemical measurements. M.T.M. and H.W.L. conducted *in situ* TEM characterization. W.Z. conducted focused ion beam experiments. N.L., Z.L. and Y.C. co-wrote the paper. All authors discussed the results and commented on the manuscript.

Additional information

Supplementary information is available in the [online version](#) of the paper. Reprints and permissions information is available online at www.nature.com/reprints. Correspondence and requests for materials should be addressed to Y.C.

Competing financial interests

The authors declare no competing financial interests.

The role of graphene for electrochemical energy storage

Rinaldo Raccichini^{1,2,3}, Alberto Varzi^{2,3}, Stefano Passerini^{2,3*} and Bruno Scrosati^{2,4*}

Since its first isolation in 2004, graphene has become one of the hottest topics in the field of materials science, and its highly appealing properties have led to a plethora of scientific papers. Among the many affected areas of materials science, this 'graphene fever' has influenced particularly the world of electrochemical energy-storage devices. Despite widespread enthusiasm, it is not yet clear whether graphene could really lead to progress in the field. Here we discuss the most recent applications of graphene — both as an active material and as an inactive component — from lithium-ion batteries and electrochemical capacitors to emerging technologies such as metal-air and magnesium-ion batteries. By critically analysing state-of-the-art technologies, we aim to address the benefits and issues of graphene-based materials, as well as outline the most promising results and applications so far.

Graphene, a carbon monolayer packed into a 2D honeycomb lattice, was for a long time considered to be merely a building block for carbonaceous materials of other dimensionalities (that is, graphite, fullerenes and carbon nanotubes)¹. Initially labelled as an 'academic material', graphene was thought not to exist in a free state until 2004, when Novoselov and co-workers isolated a single-atom-thick layer of carbon². Since then, interest in graphene has grown continuously, giving rise to what might be called the 'graphene gold rush'¹. Recently, intense research efforts — motivated by graphene's many appealing properties — have been boosted by multimillion-dollar funding from both the European Union and China³. Despite its wide range of potential applications⁴ and very promising array of features⁵ with respect to other structurally different forms of carbon (Table 1)^{5,6}, it is not yet clear whether graphene has the potential to revolutionize many aspects of our lives. In recent years, a large number of publications have discussed the application of graphene in electrochemical energy-storage devices (EESDs). However, although such discussions always highlight the advantages of graphene, they often lack an objective analysis of its limitations and drawbacks. This leaves us with a number of key questions. Will the employment of graphene be limited to niche applications, or will next-generation batteries and capacitors be graphene-based? Graphene's properties vary strongly as a function of its production method. Hence, which typologies of graphene can be produced with today's available technologies? Could these significantly outperform state-of-the-art materials? Furthermore, which performance metrics are more relevant for predicting the potential use of graphene in EESDs? This Progress Article aims to address these open questions.

Properties and production methods

Graphene — a defect-free flat carbon monolayer — is the only basic member of a much larger family of 2D carbon forms. As carefully reviewed in a *Carbon* Editorial⁷, this 'graphene family' includes materials with very different properties in terms of morphology, lateral dimensions, number of layers and defects (Tables 2 and 3)^{1,7,8}. Among these characteristics, the presence of defects is the factor that primarily affects the quality of the end material⁸ and, consequently, its electrochemical features. The methods adopted for graphene production^{5,6,9}, the most common shown in Fig. 1, play a crucial role in determining the properties of the final product.

Owing to limited scalability and high production costs, methods such as mechanical exfoliation^{2,10}, synthesis on SiC^{5,10} and bottom-up synthesis from structurally defined organic precursors^{9,10} necessarily restrict the use of graphene to fundamental research and niche applications, such as touch screens and high-frequency transistors. Similarly, chemical vapour deposition of hydrocarbons⁵, although a well-established technique in industry, seems generally unsuitable for mass-production of graphene for electrochemical energy storage because of its high cost, moderate product purity and rather low yield¹⁰. Nevertheless, chemical vapour deposition has been reported as an efficient method for producing vertically oriented graphene nanosheet electrodes¹¹, although the packing density of the as-obtained graphene is very low¹². Beyond the aforementioned techniques, two methods are widely employed for the bulk production of graphene: liquid-phase exfoliation, and reduction of graphene oxide. In liquid-phase exfoliation, pristine or expanded graphite particles, obtained by thermal expansion of graphite intercalation compounds (usually known as 'expandable graphite'), are first dispersed in a solvent to reduce the strength of the van der Waals attraction between the graphene layers. Afterwards, an external driving force such as ultrasonication¹³, electric field¹⁴ or shearing¹⁵ is used to induce the exfoliation of graphite into high-quality graphene sheets^{5,13}. Unfortunately, the low yield of this process leaves a considerable amount of unexfoliated graphite, which must be removed¹⁵. Nevertheless, the high scalability and low cost of liquid-phase exfoliation¹³ make it suitable for producing graphene in bulk quantities¹⁶. In the second method, graphene oxide (GO), a highly defective form of graphene with a disrupted sp^2 -bonding network, is produced by strong oxidation of pristine graphite^{17,18} followed by stirring or ultrasonication in liquid media¹⁹. Graphene oxide must be reduced in order to restore the π network, which is the characteristic of conductive graphene²⁰. Chemical, thermal and electrochemical processes are commonly employed in this order to produce reduced graphene oxide (RGO)^{10,20,21}. Despite the low-to-medium quality of the obtained material due to the presence of both intrinsic defects (edges and deformations) and extrinsic defects (O- and H-containing groups), these methods allow the production of bulk quantities with high yield and contained costs. Although liquid-phase exfoliation and reduction of GO are the primary methods for producing commercially available graphene for EESDs, other

¹Institute of Physical Chemistry, University of Muenster, Corrensstrasse 28/30, D-48149 Muenster, Germany. ²Helmholtz Institute Ulm, Helmholtzstrasse 11, D-89081 Ulm, Germany. ³Karlsruhe Institute of Technology, PO Box 3640, D-76021 Karlsruhe, Germany. ⁴Istituto Italiano di Tecnologia, Graphene Labs and Nanochemistry Department, Via Morego 30, I-16163 Genova, Italy. *e-mail: stefano.passerini@kit.edu; bruno.scrosati@gmail.com

Table 1 | Graphene properties compared with other carbonaceous materials.

	Graphene	Carbon nanotube	Fullerene	Graphite
Dimensions	2	1	0	3
Hybridization	sp^2	Mostly sp^2	Mostly sp^2	sp^2
Hardness	Highest (for single layer)	High	High	High
Tenacity	Flexible, elastic	Flexible, elastic	Elastic	Flexible, non-elastic
Experimental SSA ($m^2 g^{-1}$)	~1,500	~1,300	80–90	~10–20
Electrical conductivity ($S cm^{-1}$)	~2,000	Structure-dependent	10^{-10}	Anisotropic: $2-3 \times 10^{4*}$, 6^{\dagger}
Thermal conductivity ($W m^{-1} K^{-1}$)	4,840–5,300	3,500	0.4	Anisotropic: 1,500–2,000*, 5–10 [†]

*a direction, †c direction.

Table 2 | Dimension-based graphene nomenclature.

Thickness (n , number of layers)		Lateral dimension D (nm)		Aspect ratio (length:width)	
1	$2 \leq n \leq 10$	$D \leq 100$	$100 \leq D \leq 10^5$	≤ 10	> 10
Single-layer monolayer	Few-layer multilayer	Nano-	Micro-	-Sheet -Flake -Plate -Platelet	-Ribbon

Table 3 | Graphene's structural defect typologies.

Intrinsic (removal or introduction of carbon atoms in graphene's chemical composition)	Extrinsic (introduction of non-carbon atoms in graphene's chemical composition)
Vacancies	O, H and other foreign atoms
Edges	
Deformations	
Hybrid structures	

techniques are available (such as carbon nanotube unzipping²² or direct arc-discharge²³). However, owing to their higher costs, these techniques remain relatively marginal and thus unsuitable for bulk production.

In their review, Novoselov *et al.*⁵ perfectly summarized the current state of affairs: “Graphene will be of even greater interest for industrial applications when mass-produced graphene has the same outstanding performance as the best samples obtained in research laboratories.” As a matter of fact, the large-scale production of ‘outstanding performance’ graphene is the most ambitious challenge to address before its widespread application⁵. This aspect is particularly relevant with regard to the introduction of graphene in EESDs for powering millions of electric cars in the near future.

Over the past few years, many studies have explored graphene-based materials for electrochemical energy storage²⁴. In most of these, graphene was produced from graphite. As shown in Fig. 2, expandable graphite can be thermally expanded and subsequently exfoliated to obtain graphene. Pristine graphite can also be directly exfoliated to give graphene through liquid-phase methods or, alternatively, oxidized to obtain graphite oxide^{25,26}. The latter, after liquid-phase exfoliation, yields GO, which is then reduced to form RGO²⁰. This approach is different from other types of application as it is particularly useful for energy-storage materials. In fact, although oxidation introduces defects that cannot be entirely removed during the reduction process²⁰, this synthetic pathway facilitates the preparation of composites. In contrast with graphene (including RGO), GO can be easily dispersed in a wide range of solvents¹⁰. This peculiarity enables, through different chemical routes, the functionalization of GO with electroactive materials (such as conductive polymers and metal oxides) to form GO-based composites²⁷. These composites can be used as such, or alternatively can be further reduced to obtain RGO-composites²⁸.

Graphene-based materials have been proposed for use in all kinds of EESD, either as an active material or an inactive component.

Graphene as an active material

Graphene can be considered to be an active material when it takes part in an energy-storage mechanism. This can range from hosting ions (such as Li^+ or Na^+ in metal-ion batteries) to storing electrostatic charges on the electrode double-layer (as in electrochemical double-layer capacitors, EDLCs), or functioning as a catalyst in metal–air batteries.

Lithium-ion batteries. In lithium-ion batteries (LIBs), Li^+ ions continuously shuttle between a lithium-releasing cathode (commonly a layered lithium metal oxide) and a lithium-accepting anode (commonly graphite)²⁹. The amount of ions hosted per gram of material determines the capacity — and thus the energy — of the battery. Similar to graphite, graphene can be used as an anode for hosting Li^+ , both as such and as a carbonaceous matrix in composites with other materials also capable of storing lithium.

Graphene as an Li^+ host. As originally suggested by Dahn *et al.* in 1995, an anode comprising single layers of graphene can host two times as many Li^+ ions as conventional graphite^{30,31}. The storage of one lithium ion on each side of graphene results in a Li_2C_6 stoichiometry that provides a specific capacity of 744 mAh g^{-1} — twice that of graphite (372 mAh g^{-1})³⁰. This primeval concept of lithium hosting in graphene-like carbons was retrieved following the first isolation of graphene in 2004³. Differently from graphite, in which lithium is intercalated between the stacked layers³², single-layer graphene can theoretically store Li^+ ions through an adsorption mechanism, both on its internal surfaces and in the empty nanopores that exist between the randomly arranged single layers (accordingly to the ‘house of cards’ model)^{30,31}. Similarly to other disordered carbons, such a process mainly takes place at low potentials ($<0.5 \text{ V}$ versus Li/Li^+). However, it differs from the characteristic staging behaviour of graphite because graphene provides electronically and geometrically non-equivalent sites³². As a result of this unique mechanism,

the amount of lithium stored by graphene-based anodes is more strongly dependent on the production method of both the material and the electrode.

In most reported studies, RGO is the material of choice for lithium-ion storage³³. During the first Li^+ insertion, RGO exhibited incredibly high-capacity values of $>2,000 \text{ mAh g}^{-1}$ (ref. 33), which is higher than the theoretical capacity of single-layer graphene. However, this amazing capacity is not fully released after de-insertion due to the massive irreversibility of the first lithiation step³³. This phenomenon, also observed for other Li-ion anode materials^{32,34}, can be attributed mainly to the irreversible reduction of the electrolyte to form a surface passivation layer on the active particles; namely, the 'solid electrolyte interphase'³². As shown in Fig. 3a, the solid electrolyte interphase strongly depends on the specific surface area (SSA) of the active material. Accordingly, the extremely high SSA of graphene, when compared with common graphite (Table 1), results in a very high initial irreversible capacity⁶ (Fig. 3b). In the following de-insertion cycle, graphene displays a high reversible capacity, although delivered mostly at potentials of 1–3 V versus Li/Li^+ , which is rather higher than typical graphite values (0–0.4 V versus Li/Li^+). This leads to the occurrence of a large voltage hysteresis upon insertion and de-insertion of Li^+ (Fig. 3b), which results in poor energy efficiency for cells employing such electrodes. Such a drawback, together with the large cathode quantities needed to supply the initial charge for the irreversible capacity, makes graphene-based cells unfeasible. The voltage hysteresis, also observed in several nanotube-shaped materials³⁵ and high-specific-charge carbons³², is caused, among other reasons, by Li storage on defects such as edges and/or oxygen- and hydrogen-containing surface groups^{32,36}. It is thus advisable to limit the number of such defects in graphene-based anodes, particularly because they are also responsible³⁶ for the low Coulombic efficiency in the first cycle. In addition, the progressive reduction of oxygen-containing groups (for example, in RGO) leads to graphene layer re-stacking, which lowers the storage capacity over repeated cycling³³. All of these aspects affect the value of the reversible capacity, which, after a few tens of cycles, is rarely comparable to that of commercially available graphites³³.

Graphene quality is therefore a crucial issue that must be addressed before the graphite in LIBs can be replaced. Even when graphene is finally available in large quantities at reasonable cost, graphite will probably still be the active material of choice for widespread hard-case batteries, unless we develop effective strategies to prevent initial lithium ion consumption and avoid graphene layer re-stacking. In this regard, pre-lithiation^{37,38}, controlled surface functionalization⁶ and the use of composites³⁹ might be promising strategies. At the same time, the development of flexible LIBs, which require lightweight and ultrathin active materials, could benefit from the use of graphene. However, even if different studies demonstrate graphene as a promising anode in flexible LIBs, the aforementioned drawbacks still represent major obstacles for practical applications⁴⁰.

Graphene-based composite anodes. Several composites have recently been developed in an effort to overcome the energy-storage limitations and poor cycling behaviour of bare graphene negative electrodes⁴. The addition of electroactive materials, such as metal (or metal oxide) nanoparticles, provides reversible alloying (with SnO_2 or Si nanoparticles), insertion (with TiO_2) or conversion (with Fe_2O_3 or Co_3O_4) reactions with lithium, thus allowing considerably higher storage capacities than those of bare graphene or graphite^{6,41}. During the composite preparation, graphene can act as a support for the growth of electroactive nanostructures that, in turn, hinder re-stacking by lowering the van der Waals forces among the layers. As a result, graphene-based composites are less affected by agglomeration during electrode preparation, as well as by capacity fading during cycling⁶. Moreover, the extensive and highly conductive carbon matrix established by graphene layers improves the electrical

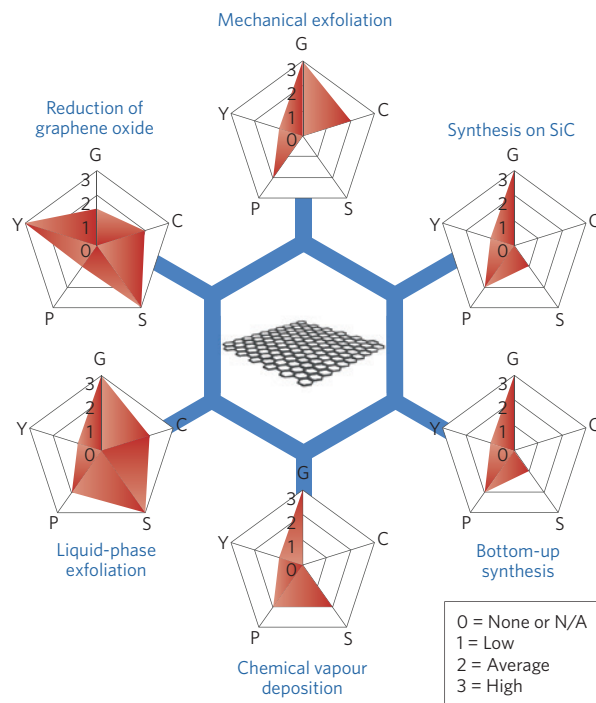


Figure 1 | Schematic of the most common graphene production methods.

Each method has been evaluated in terms of graphene quality (G), cost aspect (C; a low value corresponds to high cost of production), scalability (S), purity (P) and yield (Y) of the overall production process.

conductivity of the composite and buffers eventual volume changes taking place in electrodes based on alloying or conversion materials during cycling⁴². Despite these promising properties, however, graphene-based composites suffer, similarly to bare graphene, from the huge irreversible charge consumption of 30–50% during the first charge/discharge cycle^{6,43}. Results achieved so far with graphene composite anodes are very encouraging towards not only the development of high-energy LIBs, but also future applications such as wearable EESDs⁴⁰. Among the proposed composite graphene-based materials, some of the most promising in terms of reversible capacity are $\text{Co}_3\text{O}_4/\text{RGO}$ ($1,500 \text{ mAh g}^{-1}$)⁴⁴, silicon nanoparticles/RGO ($1,150 \text{ mAh g}^{-1}$)⁴⁴, N- and S- co-doped RGO (900 mAh g^{-1})⁴⁵ and SnO_2/RGO (700 mAh g^{-1})⁴⁶. Nevertheless, the optimization of structural arrangement and weight ratio distribution between the composite components are still key issues that must be addressed to achieve good electrochemical performance and extended cycle life⁶.

Sodium-ion batteries. The development of sodium-ion batteries (SIBs), seen as a cheaper alternative to LIBs, is promoting extensive research to identify a suitable anode active material because, owing to their large ionic radius, Na^+ ions do not intercalate into graphite⁴⁷ (Fig. 3c). In this regard, graphene seems to be a good candidate as an active anode in SIBs.

Graphene as an Na^+ host. The use of RGO as an anode material in SIBs was first reported in 2013⁴⁸, where it showed promising electrochemical behaviour, good cycle life and excellent rate capability. Such remarkable performance is related to the presence of defects (for example, residual oxygen-containing groups), which increase the graphene interlayer distance (0.37 nm, compared with 0.34 nm in graphite). However, as observed in LIBs, the presence of defects represents a serious drawback in term of Coulombic efficiency for SIBs⁴⁸. Recently, Ding *et al.* reported the synthesis of different kinds of few-layer graphene (produced from biomass precursors) and their performance as anode materials in SIBs⁴⁹. Interestingly, the obtained

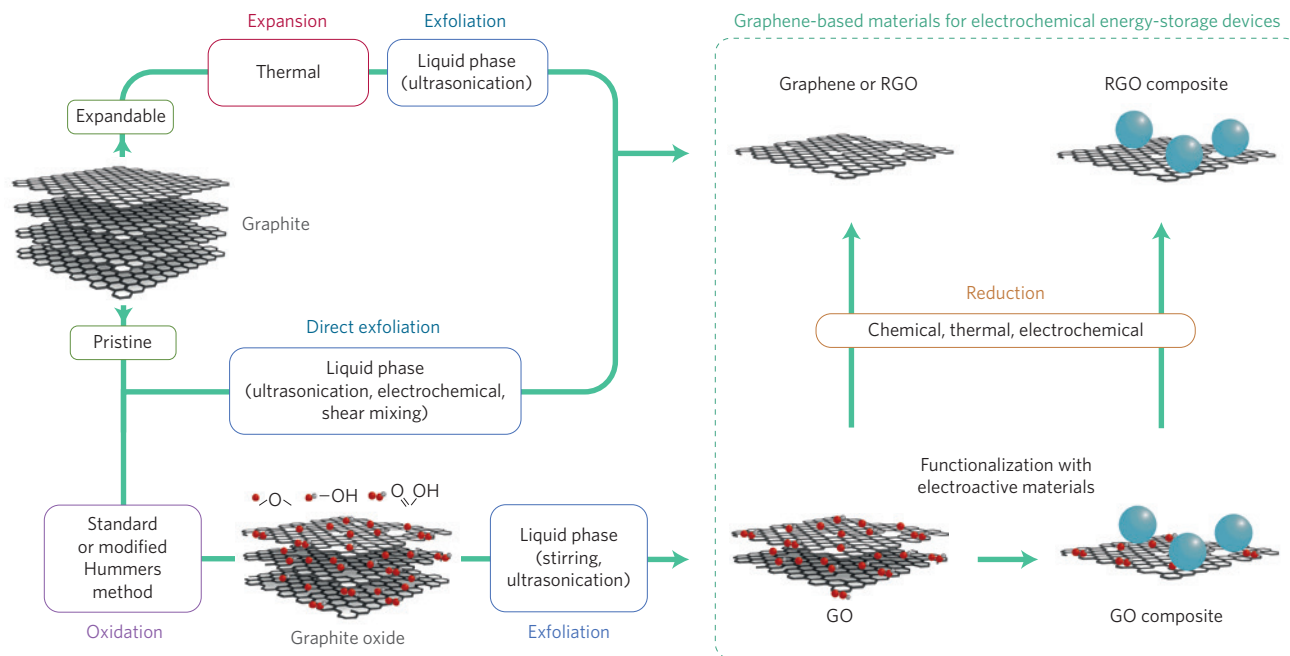


Figure 2 | The most common synthetic pathways for producing graphene-based materials (GO, RGO, GO- and RGO-based compounds) for use as electrode active materials in EESDs.

graphenes exhibited different Na^+ ion host mechanisms depending on the synthesis temperature (600–1,400 °C). Lower temperatures yielded average-quality graphenes with an Na^+ storage capacity similar to that of RGO. In contrast, higher temperatures enabled the formation of better quality graphene, with an interlayer spacing of 0.38 nm and promising insertion performance. Indeed, this report discloses one of the best-performing graphene-like materials for SIB anodes⁴⁹, showing up to 300 mAh g⁻¹ specific capacity and good retention over 200 cycles, even though the Coulombic efficiency for the first cycle remains poor. Such results give hope for the successful employment of graphene in SIBs, insofar as it can compete with other recently developed anode materials⁵⁰. Additionally, the lower insertion potential of graphene-based anodes makes it more advantageous in terms of specific energy^{49,50}.

Similarly to LIBs, graphene-based composites enable SIBs with higher specific capacity, better rate capability and longer cycle life than bare graphene^{51–55}.

Electrochemical capacitors. Electrochemical capacitors (also called supercapacitors) exploit fast charge-storage mechanisms to enable considerably higher power densities than those available in LIBs or SIBs. Electrochemical capacitors can be subdivided into two classes: electrochemical double-layer capacitors (EDLCs) and pseudocapacitors. In EDLCs, the energy is physically stored through the adsorption of ions on the surface of the electrodes, whereas in pseudocapacitors, electrochemical energy storage is enabled by fast redox reactions occurring between the electrode active material and the electrolyte⁵⁶.

Electrochemical double-layer capacitors. In EDLCs, the electrode's active materials are electrochemically stable, do not undergo any Faradaic processes and, above all, possess large SSAs⁵⁶. The amount of charge stored per unit mass (F g⁻¹), volume (F cm⁻³) or area (F cm⁻²) is indeed directly proportional to the surface available for the formation of the double layer (that is, the area in contact with the electrolyte)⁵⁷. In principle, graphene, with its theoretical SSA of 2,675 m² g⁻¹ (ref. 8) and capacitance of 550 F g⁻¹ (ref. 58), would be a perfect candidate for boosting the energy density of such devices⁵⁹. However, this does not seem to be the case in practice, as the difficulty of even approaching the theoretical SSA of graphene (for instance,

average values for RGO are in the range of 300–1,000 m² g⁻¹)⁴³ results in a lower practical gravimetric capacitance (100–270 F g⁻¹ and 70–120 F g⁻¹ with aqueous and organic electrolytes, respectively)^{43,58}. Additionally, spontaneous graphene layer re-stacking, which occurs during both electrode manufacturing and cycling, strongly reduces the practical surface available for charge storage (Fig. 3d,e). Different approaches have been introduced to mitigate these detrimental effects. As reported by Ruoff *et al.*, RGO can be chemically activated to create an extended 3D meso- and microporous network (with an SSA of up to 3,100 m² g⁻¹) of highly curved graphene walls that prevent re-stacking during cycling. Such 'activated graphene' enables high gravimetric capacitances with both organic (166 F g⁻¹) and ionic liquid electrolytes (200 F g⁻¹)⁶⁰ and, moreover, operates across a wide temperature range of -50 °C to 80 °C⁶¹. Alternatively, graphene layer re-stacking can be minimized by optimizing the electrode-manufacturing process. In this regard, RGO sheets could be vertically aligned with respect to the current collector plane, thus granting better ion accessibility and enabling higher packing density. Moreover, high and reversible volumetric (171 F cm⁻³) and areal (1.83 F cm⁻²) capacitances in aqueous electrolyte could be obtained¹².

In summary, although activated graphene and vertically aligned RGO show promising performance, the large majority of graphene-like materials cannot yet compete with the cheaper and well-established activated carbons⁶². The majority of the results reported for graphene-based supercapacitors were obtained with very low density electrode materials (for example, aerogels and foams), which possess a large number of void spaces (macropores). These pores are filled by the electrolyte, thus increasing both the weight and volume of the final device to a point where they are unsuitable for use in EDLCs^{58,63}. In contrast, graphene would probably fit in the approaching era of small-scale supercapacitors required to power the next generation of wearable- and micro-electronic devices⁶⁴.

Pseudocapacitors. In pseudocapacitors, the presence in the active material of electroactive species such as oxygen-containing functional groups, conducting polymers or transition metal oxides enables higher energy densities with respect to EDLCs^{58,59}. Nevertheless, pseudocapacitors are inferior to EDLCs in terms of power density (limited by the poor electrical conductivity of the active materials)

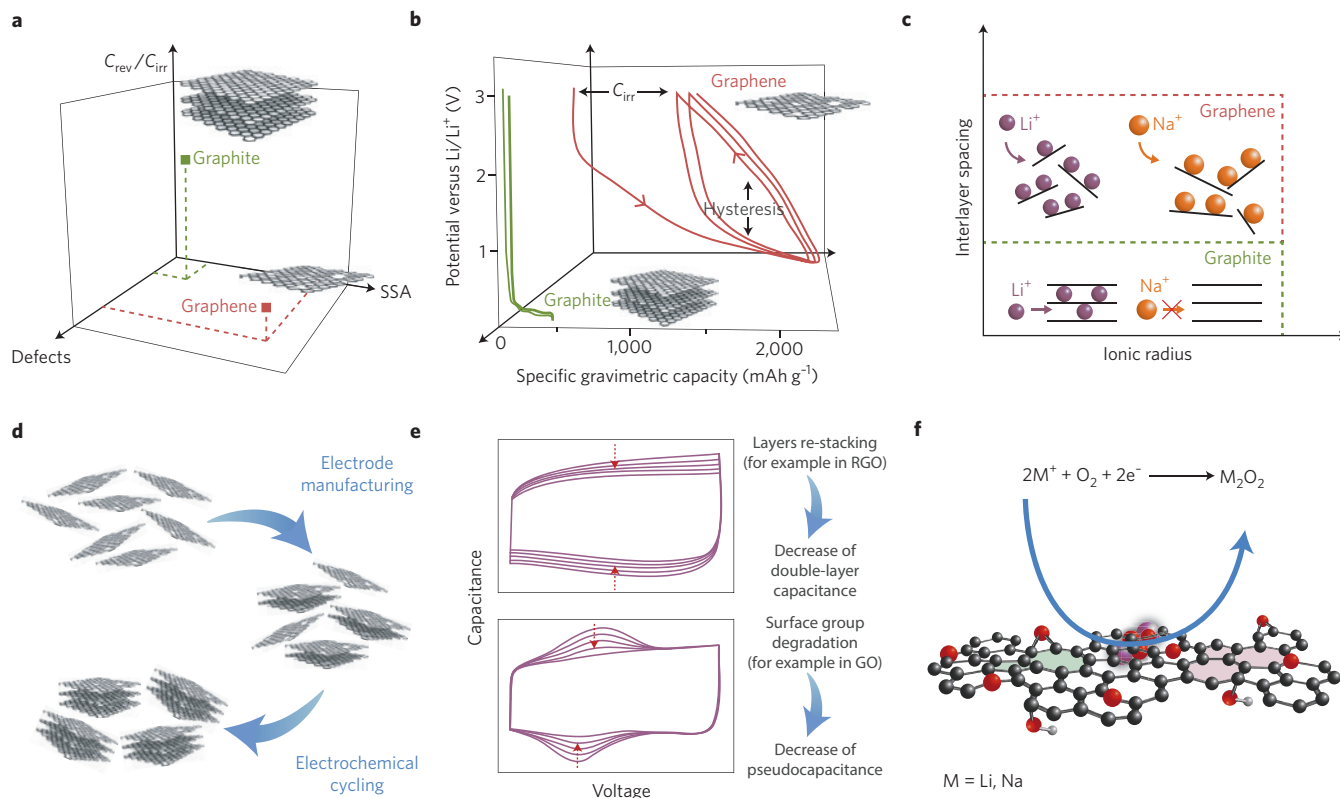


Figure 3 | Features and limitations of graphene as an active material in different EESDs. **a**, Graphite and graphene in LIB anodes. Correlation of characteristics in terms of defect amount, SSA and ratio between reversible (C_{rev}) and irreversible (C_{irr}) capacity during the first charge/discharge cycle. **b**, Typical voltage profiles of graphite and graphene (RGO) during constant current Li^+ insertion/de-insertion. **c**, Li^+ and Na^+ insertion mechanisms in graphene and graphite. **d**, Layers re-stacking in graphene during electrode manufacturing and electrochemical cycling. Re-stacking is a serious issue that affects the performances of all graphene-based EESDs. **e**, Generic voltammetric behaviour of graphene-based electrochemical capacitors over prolonged cycling. Top: Effect of graphene layers re-stacking (such as in RGO) on the double-layer capacitance. Bottom: Effect of surface group degradation (such as in GO) on pseudocapacitance. **f**, Catalytic effect of graphene defects (vacancies, deformations and presence of surface groups) in metal-air batteries.

and cycle life⁶². In this regard, graphene-based electrodes could be viable candidates for improving the performance of pseudocapacitors⁶². Despite its lower electrical conductivity, GO, owing to its large number of oxygen-containing groups, has a higher pseudocapacitance than RGO⁶⁵. However, as previously discussed, these groups may negatively affect the electrochemical behaviour of the electrode by reducing the cycling stability and reversibility^{62,65} (Fig. 3e). Various graphene-conducting polymer and graphene-metal-oxide composites have also been developed and investigated for use as pseudocapacitors^{6,62}. In these composites, graphene provides a support matrix for the growth of the electroactive species at the nanoscale, which results in a larger SSA and thus enhances the electrochemical performance by increasing electrical conductivity and mechanical stability^{62,66}.

It seems the key to exploiting the full potential of graphene in pseudocapacitors relies on the development of composite materials that offer the synergistic effect of the graphene substrate and the electroactive component, along with an optimized spatial orientation of the graphene sheets^{12,39,66}.

Lithium-air batteries. The growing demand for energy has led to the development of new EESDs with higher energy densities than metal-ion batteries. In this regard, the lithium-air battery (LAB), which offers a theoretical energy density of $5,200 \text{ Wh kg}^{-1}$ (ref. 67), represents one of the best candidates. Although lithium-air chemistry was introduced in 1976, the rechargeability of this system was brought to the attention of the scientific community only in 2006 by Bruce and colleagues⁶⁸. Although different LABs may employ

different typologies of electrolyte, they are generally composed of metallic lithium and oxygen (or air) as, respectively, the anode and cathode. The rechargeability of the system relies on the conversion of reduction products (LiO_2 and, mainly, Li_2O_2) formed during discharge (oxygen-reduction reaction), back to the original reagents during charge (oxygen-evolution reaction)³⁹. Unfortunately, the entire system suffers from a low energy efficiency, short lifetime and low rate capability (discharge capacity of 400 mAh g^{-1} after 50 cycles at a specific current of 100 mA g^{-1})^{68–70}. Reports indicate a maximum of only 100 capacity-limited ($1,000 \text{ mAh g}^{-1}$) cycles⁷¹. Among the various factors that influence the performance of LABs, the morphology of the air electrode (cathode) is particularly important for obtaining high discharge capacity. In fact, the SSA and porosity of the air electrode determine the morphology and amount of discharge products. It was demonstrated that RGO, with its large SSA, could deliver higher capacities than other carbon substrates (for example, $8,700 \text{ mAh g}^{-1}$ with respect to $1,000\text{--}2,000 \text{ mAh g}^{-1}$ in the first cycle). Defects and functional groups can also play a catalytic role for the formation of discharge products⁶⁹ (Fig. 3f). So far, the use of RGO as a bare material or substrate for other catalyst⁷² in LAB cathodes has improved performance, although achieving the theoretical energy density is still far away. Different aspects are still unclear and further studies are needed to demonstrate an effective role of graphene in LABs. Further investigations of graphene with stable electrolytes are needed before we can assess its effective role in such batteries⁶⁹.

Sodium-air batteries. Over the past five years, sodium-air batteries (SABs), despite having an energy density half that of LABs,

have been increasingly proposed owing to their low production cost and the availability of the required raw materials⁷³. In contrast with lithium, sodium is capable of reversibly forming during discharge a stable superoxide (NaO_2) with low overpotentials⁷³. This enables SABs to cycle with a charge efficiency of 80–90% after the first cycle. The formation of peroxide (Na_2O_2), however, is kinetically hindered as it requires a suitable catalyst. RGO has demonstrated, under dry air conditions, remarkable catalytic properties towards the formation of Na_2O_2 (ref. 74), which are not exhibited by conventional carbon. As reported by Liu *et al.*⁷⁴, the micro- and nanostructures of the graphene air electrode enable one of the highest specific discharge capacities for SABs. These results suggest that RGO can efficiently function as a catalyst for both oxygen-reduction and oxygen-evolution reactions. Nitrogen-doped RGO nanosheets have also been investigated in this respect. The defective sites introduced by nitrogen doping enable a more uniform and smaller size distribution of the discharge products and, therefore, a higher specific discharge capacity with respect to the undoped graphene⁷⁵. Even though this technology is in the very early stages of development, the few reports available in this field depict a quite promising scenario for graphene in SABs.

Graphene as an inactive component

Graphene could play an important role in EESDs, even without being actively involved in the electrochemical reaction. Owing to its impressive electrical conductivity (Table 1), graphene was proposed as a conductive agent in metal-ion battery electrodes as well as an encapsulating carbon matrix in, for example, lithium–sulphur batteries. Besides enabling efficient electron transport, its superior thermal conductivity (Table 1) may be advantageous for dissipating the heat generated in the case of high current loads or/and abuse conditions⁵. This would result in devices with improved intrinsic safety. The variety of structures reported can be classified into six different models (Fig. 4a)⁶.

Lithium-ion batteries. LiCoO_2 , LiMn_2O_4 and LiFePO_4 (hereafter referred to as LMO) are some of the most commonly used cathode materials in LIBs. The cycle life and rate capability of these materials are generally limited by their poor electrical conductivity. Introducing low-cost conductive additives (for example, carbon black) into the composite electrodes commonly solves this issue. Nevertheless, owing to their amorphous structure, carbon blacks possess a rather low electrical conductivity, with respect to more crystalline carbons such as graphene^{5,76}.

Recently, Kucinskis *et al.*⁷⁶ reviewed state-of-the-art graphene-based composite cathode materials. Among the vast number of reports, most employ GO as a source for the formation of graphene conductive networks. Additionally, in a large part of these works, GO is reduced to RGO simultaneously with the LMO precursors (one-pot synthesis) to produce graphene-based composites. This approach, which is different from simply mixing the carbon additive with the LMO active material during electrode preparation, promotes the formation of small-size LMO particles (leading to improved Li^+ diffusion) directly onto the RGO matrix⁷⁶. Moreover, the RGO 3D network is reportedly capable of preventing the dissolution of some LMOs³⁹, thus extending the cycle life of the batteries. However, it was also suggested that when RGO is mixed in a manner similar to conventional carbon additives during electrode preparation, it could negatively affect the Li^+ mobility, thus worsening the electrochemical performance of the composite cathode⁷⁶ (Fig. 4b). Regardless of this fact, RGO is generally reported to enhance the rate capability of the cathode with respect to conventional carbon additives. Depending on the active material, improvements of up to 160% of the discharge capacity (at the same current rate) have been observed⁷⁶. Nevertheless, it is not yet clear whether RGO may replace carbon blacks, which are much cheaper and easier to handle⁷⁶.

Sodium-ion batteries. As explained previously for the negative electrode, the larger size of Na^+ with respect to Li^+ restricts the choice of available cathode material for SIBs⁴⁷. Several layered oxides with promising electrochemical performance have recently been developed⁵⁰. However, like their LIB analogues, they usually possess poor electrical conductivity and thus limited rate capability. So far, only a few reports are available on graphene-containing composite cathodes for SIBs^{77–80}. However, in all cases, the RGO matrix seems to enhance the electrical conductivity of the composite, thus improving the rate capability compared with bare cathode materials^{77–80}.

Lithium–sulphur batteries. Lithium–sulphur batteries (LSBs), through the redox reaction of metallic lithium (anode) and elemental sulphur (cathode), could provide a remarkably high theoretical specific energy of up to 2,600 Wh kg^{-1} (ref. 39). Despite the intrinsic advantages of sulphur in terms of low cost, abundance, low toxicity and high theoretical specific capacity (1,672 mAh g^{-1}), LSBs are affected by several drawbacks: (1) slow kinetics owing to the low electrical conductivity of the redox reaction products; (2) low energy efficiency; (3) poor cycle life as a direct result of the dissolution of the intermediate reaction products (polysulphides) in the electrolyte; and (4) large volume changes during the electrochemical reaction⁶⁹. Graphene has been proposed as a good candidate to address these issues because of its high electrical conductivity and capability of trapping the charge/discharge products³⁹. Several studies have reported that RGO and GO are suitable substrates for the deposition of sulphur micro- and nanoparticles^{39,69}. Good encapsulation seems to be achieved in both cases; however, the presence of epoxy and hydroxyl groups in GO promotes the immobilization of sulphur and thus prevents its dissolution³⁹. Promising performance — with specific capacities $>1,000$ mAh g^{-1} — has also been obtained with hybrid graphene–polymer–sulphur composites, although an acceptable capacity retention during cycling has not yet been achieved⁴¹. Recently, different graphene–sulphur composites have been synthesized and tested in LSBs^{81–83}. They exhibit a good performance in terms of capacity, Coulombic efficiency and stability during cycling^{81–84}, even if the values reported are not higher than those obtained with other carbon-encapsulated sulphur cathodes, such as ordered mesoporous carbon⁸⁴.

From these results, graphene might be a possible candidate for encapsulating sulphur on LSB cathodes. However, real advances in the field require improvements to the sulphur/graphene interface in order to achieve stable electrochemical performance³⁹.

Developing applications of graphene

The recent outbreak of graphene in the field of electrochemical energy storage has spurred research into its applications in novel systems such as magnesium-ion batteries (MIBs), which is one of the newest members of the metal-ion battery family. MIBs have been proposed as a high-energy-density and environmentally friendly replacement for LIBs⁸⁵. Although research in this field is still at an early stage, a few graphene-based composites have recently been proposed as MIB cathode materials^{86,87}. Although the results obtained so far show very poor electrochemical behaviour, they could represent the first attempts to use graphene in MIBs. Interestingly, Wang *et al.* have already patented a rechargeable magnesium-ion cell based on graphene active materials⁸⁸.

The employment of graphene has also been considered for the improvement of vanadium redox-flow batteries (VRFBs). Patented in 1986⁸⁹, VRFBs enable energy storage using $\text{V}^{3+}/\text{V}^{2+}$ and $[\text{VO}_2^+]/[\text{VO}_2^{2+}]$ redox couples as negative and positive acid electrolytic solutions, respectively⁹⁰. Carbon-based materials, such as cloths or graphite felts, are used as electrodes because of their electrochemical stability and wide operating potential⁹¹. Unfortunately, despite their high SSAs, these electrodes do not exhibit satisfactory electrochemical properties⁹². Accordingly, the use of graphene-based materials has

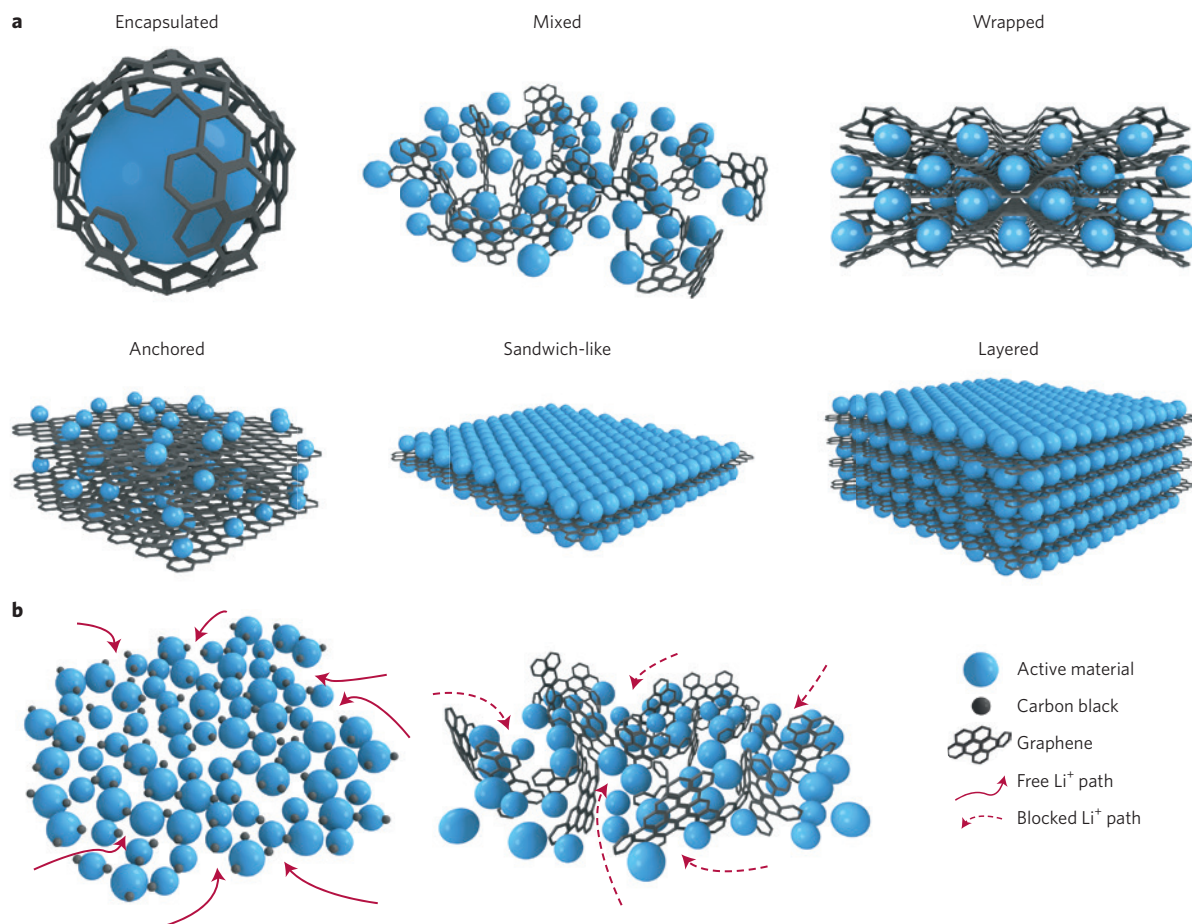


Figure 4 | Structural models and a possible drawback of graphene composites. **a**, Schematic of the different structures of graphene composite electrode materials. All models (except where specifically indicated) refer to composites in which graphene and the active material are synthesized through one-pot processes. Encapsulated: Single active-material particles are encapsulated by graphene, which acts as either an active (for example, LIB anodes) or an inactive (for example, LIB cathodes) component. Mixed: Graphene and active materials are synthesized separately and mixed mechanically during the electrode preparation. In this structure, graphene may serve as an inactive conductive matrix (for example, LIB cathodes) or an active material (for example, LIB anodes). Wrapped: The active-material particles are wrapped by multiple graphene sheets. This structure well-represents pseudocapacitor electrodes, in which graphene is the active material, as well as metal-ion battery cathodes, where graphene is an inactive component. Anchored: This is the most common structure for graphene composites, in which electroactive nanoparticles are anchored to the graphene surface. This structure is very relevant for metal-ion battery anodes and pseudocapacitors, where graphene serves as an active material, as well as for metal-ion battery cathodes and in LSBs, where graphene acts as an inactive component. Sandwich-like model: Graphene is used as a template to generate active material/graphene sandwich structures. This graphene-composite model, although not widespread, is used for LIB cathodes. Layered model: Active-material nanoparticles are alternated with graphene sheets to form a composite layered structure, which has been proposed for use in metal-ion battery anodes and cathodes. **b**, Li^+ paths in carbon black- (left) and graphene- (right) based electrodes in the mixed structural model. The figure highlights a possible drawback of graphene in terms of Li^+ mobility.

been proposed to improve electrical conductivity, kinetic reversibility and electrochemical activity of these electrodes^{91,92}. Over the past few years, a small number of studies have investigated the electrochemical properties of GO-based⁹¹, RGO^{90,92} and RGO-based composites^{93–95} for application in VRFBs. All of these reports show promising electrochemical performance for graphene-supported carbon electrodes, specifically in terms of high peak current density, reduced overpotential and decreased charge-transfer resistance. Additionally, GO⁹⁶ and commercial graphene⁹⁷ have recently been tested as additives in VRFB ion-exchange membranes, with the aim of reducing the vanadium permeation and preventing ionic cross-mixing. The results achieved so far seem promising when compared with those obtained with bare membranes. However, the development of a successful commercial VRFB containing graphene is still far away.

Conclusions

It has been ten years since the beginning of the graphene era, and the rush to find new applications for this exciting material is more

vibrant than ever. However, despite the enormous amount of data produced throughout research laboratories across the globe, it is still not clear whether graphene has the potential to revolutionize many aspects of our lives. This is particularly appropriate for the field of electrochemical energy storage, in which ‘graphene fever’ has reached rather high levels due to the continuous need for new materials that can meet the market’s performance requirements.

Graphene promises to increase substantially the energy- and power-density of practical systems, as well as enable the development of next-generation devices. However, the results so far tend to suggest that real breakthroughs are still to come. As was the case for many other innovative materials in the past, the main task is to close the gap between laboratory-scale research and practical applications. The first challenge lies in the production of graphene. Owing to its peculiar nature, the electrochemical properties of this material are strictly dependent on its method of production, and so are its chances of finding an application in EESDs. Nowadays, the vast majority of graphene-based materials are produced by the reduction

of GO to RGO. This method is relatively cheap and has the potential for scalability — mandatory properties for widespread adoption — but introduces both intrinsic and extrinsic defects, which strongly affect electrochemical properties. Although for some applications the defects function as catalyst sites to improve cell kinetics (for example, in lithium–air and sodium–air batteries), in other cases they strongly reduce the performance. For example, RGO cannot compete with the widespread carbonaceous materials commonly employed in commercial LIBs. In fact, despite their very promising initial performance, RGO electrodes show a limited cycle life⁹⁸ compared with well-established graphite electrodes. With respect to EDLCs, the limited cycle life and low density of RGO-based electrodes prevents their transition to the commercial stage. However, some strategies for improving the packing density of graphene-based materials have been proposed. Nevertheless, the macroporous nature of graphene, in general, seriously affects its volumetric energy density. In this case, the common practice of evaluating EDLC performance through gravimetric data might lead to misleading conclusions. Because low-density and few-micrometre-thick electrodes are often reported, volumetric data are surely more appropriate⁶³.

In view of the funding and human resources devoted worldwide to this unique material, we may expect to see a turnover in the not-too-distant future. Some important results support this vision. In fact, a growing body of research into graphene-based full LIBs^{37,38,99,100} is continuing to prove the benefits of graphene for this important application. In addition, it has been demonstrated that graphene (or RGO) may find its true role when employed in composite electrodes. Here, graphene layers and electroactive particles work symbiotically, with the former providing a stiff and conductive matrix, which can buffer eventual volume changes, and the latter helping to avoid layer re-stacking. Graphene-based composites have, in fact, shown outstanding performance in LIBs^{44–46}, SIBs^{51–55}, pseudocapacitors^{58,59,62,66} and LSB^{81–83}. Moreover, a few preliminary studies into full SIBs^{51,55} have confirmed graphene exploitation in the ‘beyond lithium’ battery generation. Nevertheless, the most crucial point is the nano-architecture of the composite. Indeed, if proper nanoscale engineering is achieved, such compounds will surely play a crucial role in the progress of the field.

Winning the graphene ‘gold rush’ requires consistent investment and commitment from industry and research-funding institutions. In this scenario, research scientists are those who occupy the most prominent position, by highlighting the benefits and, most importantly, addressing the issues that still hinder the large-scale application of graphene in EESDs.

Received 4 March 2014; accepted 7 November 2014; published online 22 December 2014

References

- Geim, A. K. & Novoselov, K. S. The rise of graphene. *Nature Mater.* **6**, 183–191 (2007).
- Novoselov, K. S. *et al.* Electric field effect in atomically thin carbon films. *Science* **306**, 666–669 (2004).
- Graphene Flagship; <http://graphene-flagship.eu/>.
- Singh, V. *et al.* Graphene based materials: Past, present and future. *Prog. Mater. Sci.* **56**, 1178–1271 (2011).
- Novoselov, K. S. *et al.* A roadmap for graphene. *Nature* **490**, 192–200 (2012).
- Wu, Z.-S. *et al.* Graphene/metal oxide composite electrode materials for energy storage. *Nano Energy* **1**, 107–131 (2012).
- Bianco, A. *et al.* All in the graphene family – A recommended nomenclature for two-dimensional carbon materials. *Carbon* **65**, 1–6 (2013).
- Ivanovskii, A. L. Graphene-based and graphene-like materials. *Russ. Chem. Rev.* **81**, 571–605 (2012).
- Sivudu, K. S. & Mahajan, Y. R. Challenges and opportunities for the mass production of high quality graphene: An analysis of worldwide patents. *Nanotech Insights* **3**, 6–18 (2012).
- Warner, J. H., Schäffel, F., Bachmatiuk, A. & Rummeli, M. H. *Graphene: Fundamentals and Emergent Applications* Ch. 4 (Elsevier, 2013).
- Miller, J. R., Outlaw, R. A. & Holloway, B. C. Graphene double-layer capacitor with ac line-filtering performance. *Science* **329**, 1637–1639 (2010).
- Yoon, Y. *et al.* Vertical alignments of graphene sheets spatially and densely piled for fast ion diffusion in compact supercapacitors. *ACS Nano* **8**, 4580–4590 (2014).
- Cai, M., Thorpe, D., Adamson, D. H. & Schniepp, H. C. Methods of graphite exfoliation. *J. Mater. Chem.* **22**, 24992–25002 (2012).
- Wei, D. *et al.* Graphene from electrochemical exfoliation and its direct applications in enhanced energy storage devices. *Chem. Commun.* **48**, 1239–1241 (2012).
- Paton, K. R. *et al.* Scalable production of large quantities of defect-free few-layer graphene by shear exfoliation in liquids. *Nature Mater.* **13**, 624–630 (2014).
- Tour, J. M. Layered materials: Scaling up exfoliation. *Nature Mater.* **13**, 545–546 (2014).
- Hummers, W. S. Jr & Offeman, R. E. Preparation of graphitic oxide. *J. Am. Chem. Soc.* **80**, 1339 (1957).
- Kovtyukhova, N. I. *et al.* Layer-by-layer assembly of ultrathin composite films from micron-sized graphite oxide sheets and polycations. *Chem. Mater.* **11**, 771–778 (1999).
- Li, D., Müller, M. B., Gilje, S., Kaner, R. B. & Wallace, G. G. Processable aqueous dispersions of graphene nanosheets. *Nature Nanotech.* **3**, 101–105 (2008).
- Dreyer, D. R., Park, S., Bielawski, C. W. & Ruoff, R. S. The chemistry of graphene oxide. *Chem. Soc. Rev.* **39**, 228–240 (2010).
- Park, S. & Ruoff, R. S. Chemical methods for the production of graphenes. *Nature Nanotech.* **4**, 217–224 (2009).
- Kosynkin, D. V. *et al.* Longitudinal unzipping of carbon nanotubes to form graphene nanoribbons. *Nature* **458**, 872–876 (2009).
- Wu, Y. *et al.* Efficient and large-scale synthesis of few-layered graphene using an arc-discharge method and conductivity studies of the resulting films. *Nano Res.* **3**, 661–669 (2010).
- Hou, J., Shao, Y., Ellis, M. W., Moore, R. B. & Yi, B. Graphene-based electrochemical energy conversion and storage: Fuel cells, supercapacitors and lithium ion batteries. *Phys. Chem. Chem. Phys.* **13**, 15384–15402 (2011).
- Hirata, M., Gotou, T., Horiuchi, S., Fujiwara, M. & Ohba, M. Thin-film particles of graphite oxide 1: high-yield synthesis and flexibility of the particles. *Carbon* **42**, 2929–2937 (2004).
- Gao, W., Alemany, L. B., Ci, L. & Ajayan, P. M. New insights into the structure and reduction of graphite oxide. *Nature Chem.* **1**, 403–408 (2009).
- Compton, O. C. & Nguyen, S. T. Graphene oxide, highly reduced graphene oxide, and graphene: Versatile building blocks for carbon-based materials. *Small* **6**, 711–723 (2010).
- Stankovich, S. *et al.* Graphene-based composite materials. *Nature* **442**, 282–286 (2006).
- Scrosati, B. & Garche, J. Lithium batteries: status, prospects and future. *J. Power Sources* **195**, 2419–2430 (2010).
- Dahn, J. R., Zheng, T., Liu, Y. & Xue, J. S. Mechanisms for lithium insertion in carbonaceous materials. *Science* **270**, 590–593 (1995).
- Liu, Y., Xue, J. S., Zheng, T. & Dahn, J. R. Mechanism of lithium insertion in hard carbons prepared by pyrolysis of epoxy resins. *Carbon* **34**, 193–200 (1996).
- Winter, M., Besenhard, J. O., Spahr, M. E. & Novák, P. Insertion electrode materials for rechargeable lithium batteries. *Adv. Mater.* **10**, 725–763 (1998).
- Vargas, C. O. A., Caballero, Á. & Morales, J. Can the performance of graphene nanosheets for lithium storage in Li-ion batteries be predicted? *Nanoscale* **4**, 2083–2092 (2012).
- Zhang, W.-J. A review of the electrochemical performance of alloy anodes for lithium-ion batteries. *J. Power Sources* **196**, 13–24 (2011).
- Landi, B. J., Ganter, M. J., Cress, C. D., DiLeo, R. A. & Raffaele, R. P. Carbon nanotubes for lithium ion batteries. *Energ. Environ. Sci.* **2**, 638–654 (2009).
- Xiang, H. F. *et al.* Graphene sheets as anode materials for Li-ion batteries: Preparation, structure, electrochemical properties and mechanism for lithium storage. *RSC Adv.* **2**, 6792–6799 (2012).
- Vargas, O., Caballero, Á. & Morales, J. Enhanced electrochemical performance of maghemite/graphene nanosheets composite as electrode in half and full Li-ion cells. *Electrochim. Acta* **130**, 551–558 (2014).
- Hassoun, J. *et al.* An advanced lithium-ion battery based on a graphene anode and a lithium iron phosphate cathode. *Nano Lett.* **14**, 4901–4906 (2014).
- Zhu, J., Yang, D., Yin, Z., Yan, Q. & Zhang, H. Graphene and graphene-based materials for energy storage applications. *Small* **10**, 3480–3498 (2014).
- Zhou, G., Li, F. & Cheng, H.-M. Progress in flexible lithium batteries and future prospects. *Energ. Environ. Sci.* **7**, 1307–1338 (2014).
- Xu, C. *et al.* Graphene-based electrodes for electrochemical energy storage. *Energ. Environ. Sci.* **6**, 1388–1414 (2013).
- Huang, X., Zeng, Z., Fan, Z., Liu, J. & Zhang, H. Graphene-based electrodes. *Adv. Mater.* **24**, 5979–6004 (2012).
- Sun, Y., Wu, Q. & Shi, G. Graphene based new energy materials. *Energ. Environ. Sci.* **4**, 1113–1132 (2011).
- Lee, W. W. & Lee, J.-M. Novel synthesis of high performance anode materials for lithium-ion batteries (LIBs). *J. Mater. Chem. A* **2**, 1589–1626 (2014).

45. Ai, W. *et al.* Nitrogen and sulfur codoped graphene: Multifunctional electrode materials for high-performance Li-ion batteries and oxygen reduction reaction. *Adv. Mater.* **26**, 6186–6192 (2014).
46. Chen, J. S. & Lou, X. W. D. SnO₂-based nanomaterials: Synthesis and application in lithium-ion batteries. *Small* **9**, 1877–1893 (2013).
47. Cao, Y. *et al.* Sodium ion insertion in hollow carbon nanowires for battery applications. *Nano Lett.* **12**, 3783–3787 (2012).
48. Wang, Y.-X., Chou, S.-L., Liu, H.-K. & Dou, S.-X. Reduced graphene oxide with superior cycling stability and rate capability for sodium storage. *Carbon* **57**, 202–208 (2013).
49. Ding, J. *et al.* Carbon nanosheet frameworks derived from peat moss as high performance sodium ion battery anodes. *ACS Nano* **7**, 11004–11015 (2013).
50. Hong, S. Y. *et al.* Charge carriers in rechargeable batteries: Na ions vs. Li ions. *Energ. Environ. Sci.* **6**, 2067–2081 (2013).
51. Yu, D. Y. W. *et al.* High-capacity antimony sulphide nanoparticle-decorated graphene composite as anode for sodium-ion batteries. *Nature Commun.* **4**, 2922 (2013).
52. Prihodchenko, P. V. *et al.* Nanocrystalline tin disulfide coating of reduced graphene oxide produced by the peroxostannate deposition route for sodium ion battery anodes. *J. Mater. Chem. A* **2**, 8431–8437 (2014).
53. Nithya, C. & Gopukumar, S. rGO/nano Sb composite: A high performance anode material for Na⁺ ion batteries and evidence for the formation of nanoribbons from the nano rGO sheet during galvanostatic cycling. *J. Mater. Chem. A* **2**, 10516–10525 (2014).
54. Qin, G., Zhang, X. & Wang, C. Design of nitrogen doped graphene grafted TiO₂ hollow nanostructures with enhanced sodium storage performance. *J. Mater. Chem. A* **2**, 12449–12458 (2014).
55. Qu, B. *et al.* Layered SnS₂-reduced graphene oxide composite — A high-capacity, high-rate, and long-cycle life sodium-ion battery anode material. *Adv. Mater.* **26**, 3854–3859 (2014).
56. Simon, P. & Gogotsi, Y. Materials for electrochemical capacitors. *Nature Mater.* **7**, 845–854 (2008).
57. Stoller, M. D., Park, S., Zhu, Y., An, J. & Ruoff, R. S. Graphene-based ultracapacitors. *Nano Lett.* **8**, 3498–3502 (2008).
58. Chen, J., Li, C. & Shi, G. Graphene materials for electrochemical capacitors. *J. Phys. Chem. Lett.* **4**, 1244–1253 (2013).
59. Bose, S. *et al.* Carbon-based nanostructured materials and their composites as supercapacitor electrodes. *J. Mater. Chem.* **22**, 767–784 (2012).
60. Zhu, Y. *et al.* Carbon-based supercapacitors produced by activation of graphene. *Science* **332**, 1537–1541 (2011).
61. Tsai, W.-Y. *et al.* Outstanding performance of activated graphene based supercapacitors in ionic liquid electrolyte from –50 to 80°C. *Nano Energy* **2**, 403–411 (2013).
62. Huang, Y., Liang, J. & Chen, Y. An overview of the applications of graphene-based materials in supercapacitors. *Small* **8**, 1805–1834 (2012).
63. Gogotsi, Y. & Simon, P. True performance metrics in electrochemical energy storage. *Science* **334**, 917–918 (2011).
64. Beidaghi, M. & Gogotsi, Y. Capacitive energy storage in micro-scale devices: Recent advances in design and fabrication of micro-supercapacitors. *Energ. Environ. Sci.* **7**, 867–884 (2014).
65. Xu, B. *et al.* What is the choice for supercapacitors: Graphene or graphene oxide? *Energ. Environ. Sci.* **4**, 2826–2830 (2011).
66. Han, G. *et al.* MnO₂ nanorods intercalating graphene oxide/polyaniline ternary composites for robust high-performance supercapacitors. *Sci. Rep.* **4**, 4824 (2014).
67. Lee, J.-S. *et al.* Metal–air batteries with high energy density: Li–air versus Zn–air. *Adv. Energ. Mater.* **1**, 34–50 (2011).
68. Ogasawara, T., Débart, A., Holzapfel, M., Novák, P. & Bruce, P. G. Rechargeable Li₂O₂ electrode for lithium batteries. *J. Am. Chem. Soc.* **128**, 1390–1393 (2006).
69. Kim, H., Lim, H.-D., Kim, J. & Kang, K. Graphene for advanced Li/S and Li/air batteries. *J. Mater. Chem. A* **2**, 33–47 (2014).
70. Girishkumar, G., McCloskey, B., Luntz, A. C., Swanson, S. & Wilcke, W. Lithium–air battery: Promise and challenges. *J. Phys. Chem. Lett.* **1**, 2193–2203 (2010).
71. Jung, H.-G., Hassoun, J., Park, J.-B., Sun, Y.-K. & Scrosati, B. An improved high-performance lithium–air battery. *Nature Chem.* **4**, 579–585 (2012).
72. Jung, H.-G. *et al.* Ruthenium-based electrocatalysts supported on reduced graphene oxide for lithium–air batteries. *ACS Nano* **7**, 3532–3539 (2013).
73. Hartmann, P. *et al.* A rechargeable room-temperature sodium superoxide (NaO₂) battery. *Nature Mater.* **12**, 228–232 (2013).
74. Liu, W., Sun, Q., Yang, Y., Xie, J.-Y. & Fu, Z.-W. An enhanced electrochemical performance of a sodium–air battery with graphene nanosheets as air electrode catalysts. *Chem. Commun.* **49**, 1951–1953 (2013).
75. Li, Y. *et al.* Superior catalytic activity of nitrogen-doped graphene cathodes for high energy capacity sodium–air batteries. *Chem. Commun.* **49**, 11731–11733 (2013).
76. Kucinskis, G., Bajars, G. & Kleperis, J. Graphene in lithium ion battery cathode materials: A review. *J. Power Sources* **240**, 66–79 (2013).
77. Jung, Y. H., Lim, C. H. & Kim, D. K. Graphene-supported Na₃V₂(PO₄)₃ as a high rate cathode material for sodium–ion batteries. *J. Mater. Chem. A* **1**, 11350–11354 (2013).
78. Xu, M. *et al.* Na₃V₂O₂(PO₄)₂/F/graphene sandwich structure for high-performance cathode of a sodium–ion battery. *Phys. Chem. Chem. Phys.* **15**, 13032–13037 (2013).
79. Zhu, H. *et al.* Free-standing Na_{2/3}Fe_{1/2}Mn_{1/2}O₂@graphene film for a sodium–ion battery cathode. *ACS Appl. Mater. Interfaces* **6**, 4242–4247 (2014).
80. Yang, D., Liao, X.-Z., Shen, J., He, Y.-S. & Ma, Z.-F. A flexible and binder-free reduced graphene oxide/Na_{2/3}Ni_{1/3}Mn_{2/3}O₂ composite electrode for high-performance sodium ion batteries. *J. Mater. Chem. A* **2**, 6723–6726 (2014).
81. Zu, C. & Manthiram, A. Hydroxylated graphene-sulfur nanocomposites for high-rate lithium–sulfur batteries. *Adv. Energ. Mater.* **3**, 1008–1012 (2013).
82. Zhao, M.-Q. *et al.* Unstacked double-layer templated graphene for high-rate lithium–sulfur batteries. *Nature Commun.* **5**, 3410 (2014).
83. Lu, S., Chen, Y., Wu, X., Wang, Z. & Li, Y. Three-dimensional sulfur/graphene multifunctional hybrid sponges for lithium–sulfur batteries with large areal mass loading. *Sci. Rep.* **4**, 4629 (2014).
84. Yin, Y.-X., Xin, S., Guo, Y.-G. & Wan, L.-J. Lithium–sulfur batteries: Electrochemistry, materials, and prospects. *Angew. Chem. Int. Ed.* **52**, 13186–13200 (2013).
85. Yoo, H. D. *et al.* Mg rechargeable batteries: An on-going challenge. *Energ. Environ. Sci.* **6**, 2265–2279 (2013).
86. Liu, Y. *et al.* Synthesis of rGO-supported layered MoS₂ for high-performance rechargeable Mg batteries. *Nanoscale* **5**, 9562–9567 (2013).
87. Chen, Q. *et al.* PTMA/graphene as a novel cathode material for rechargeable magnesium batteries. *Acta Physico-Chimica Sin.* **29**, 2295–2299 (2013).
88. Wang, Y., Zhamu, A. & Jang, B. Z. Rechargeable magnesium-ion cell having a high-capacity cathode. US Patent 2013/0302697 (2013).
89. Weber, A. Z. *et al.* Redox flow batteries: A review. *J. Appl. Electrochem.* **41**, 1137–1164 (2011).
90. González, Z. *et al.* Graphite oxide-based graphene materials as positive electrodes in vanadium redox flow batteries. *J. Power Sources* **241**, 349–354 (2013).
91. Han, P. *et al.* Graphene oxide nanosheets/multi-walled carbon nanotubes hybrid as an excellent electrocatalytic material towards VO²⁺/VO³⁺ redox couples for vanadium redox flow batteries. *Energ. Environ. Sci.* **4**, 4710–4717 (2011).
92. González, Z. *et al.* Thermally reduced graphite oxide as positive electrode in vanadium redox flow batteries. *Carbon* **50**, 828–834 (2012).
93. Flox, C., Skoumal, M., Rubio-García, J., Andreu, T. & Morante, J. R. Strategies for enhancing electrochemical activity of carbon-based electrodes for all-vanadium redox flow batteries. *Appl. Energy* **109**, 344–351 (2013).
94. Han, P. *et al.* RuSe/reduced graphene oxide: An efficient electrocatalyst for VO²⁺/VO³⁺ redox couples in vanadium redox flow batteries. *RSC Adv.* **4**, 20379–20381 (2014).
95. Shi, L., Liu, S., He, Z. & Shen, J. Nitrogen-doped graphene: Effects of nitrogen species on the properties of the vanadium redox flow battery. *Electrochim. Acta* **138**, 93–100 (2014).
96. Dai, W., Shen, Y., Li, Z., Yu, L. & Qiu, X. SPEEK/graphene oxide nanocomposite membranes with superior cyclability for highly efficient vanadium redox flow battery. *J. Mater. Chem. A* **2**, 12423–12432 (2014).
97. Dai, W. *et al.* Sulfonated poly(ether ether ketone)/graphene composite membrane for vanadium redox flow battery. *Electrochim. Acta* **132**, 200–207 (2014).
98. Vargas, O. *et al.* Electrochemical performance of a graphene nanosheets anode in a high voltage lithium-ion cell. *Phys. Chem. Chem. Phys.* **15**, 20444–20446 (2013).
99. Choi, D. *et al.* Li-ion batteries from LiFePO₄ cathode and anatase/graphene composite anode for stationary energy storage. *Electrochem. Commun.* **12**, 378–381 (2010).
100. Chae, C., Noh, H.-J., Lee, J. K., Scrosati, B. & Sun, Y.-K. A high-energy Li-ion battery using a silicon-based anode and a nano-structured layered composite cathode. *Adv. Func. Mater.* **24**, 3036–3042 (2014).

Acknowledgements

R.R., A.V. and S.P. acknowledge the financial support of Bundesministerium für Bildung und Forschung (BMBF) within the project 'IES, Innovative Elektrochemische Superkondensatoren' (contract number 03EK3010). B.S. is grateful to the Helmholtz Institute Ulm for a six-month visiting professorship position.

Author contributions

R.R. and A.V. designed the outline of the Progress Article, wrote the manuscript and conceived the figures and tables. S.P. and B.S. supervised and revised the writing.

Additional information

Reprints and permissions information is available online at www.nature.com/reprints. Correspondence should be addressed to S.P. or B.S.

Competing financial interests

The authors declare no competing financial interests.

Interconnected hollow carbon nanospheres for stable lithium metal anodes

Guangyuan Zheng¹, Seok Woo Lee², Zheng Liang², Hyun-Wook Lee², Kai Yan², Hongbin Yao², Haotian Wang³, Weiyang Li², Steven Chu⁴ and Yi Cui^{2,5*}

For future applications in portable electronics, electric vehicles and grid storage, batteries with higher energy storage density than existing lithium ion batteries need to be developed. Recent efforts in this direction have focused on high-capacity electrode materials such as lithium metal, silicon and tin as anodes, and sulphur and oxygen as cathodes. Lithium metal would be the optimal choice as an anode material, because it has the highest specific capacity (3,860 mAh g⁻¹) and the lowest anode potential of all. However, the lithium anode forms dendritic and mossy metal deposits, leading to serious safety concerns and low Coulombic efficiency during charge/discharge cycles. Although advanced characterization techniques have helped shed light on the lithium growth process, effective strategies to improve lithium metal anode cycling remain elusive. Here, we show that coating the lithium metal anode with a monolayer of interconnected amorphous hollow carbon nanospheres helps isolate the lithium metal depositions and facilitates the formation of a stable solid electrolyte interphase. We show that lithium dendrites do not form up to a practical current density of 1 mA cm⁻². The Coulombic efficiency improves to ~99% for more than 150 cycles. This is significantly better than the bare unmodified samples, which usually show rapid Coulombic efficiency decay in fewer than 100 cycles. Our results indicate that nanoscale interfacial engineering could be a promising strategy to tackle the intrinsic problems of lithium metal anodes.

When interest in secondary lithium batteries began to emerge more than four decades ago¹ it was clear that, to make viable Li metal anodes, two fundamental challenges would need to be resolved: (1) accommodating the large change in electrode volume during cycling (unlike graphite and silicon anodes, where lithiation produces volume changes of ~10% and 400%, respectively, Li metal is 'hostless' and its relative volumetric change is virtually infinite); and (2) controlling the reactivity towards the electrolyte (Li is one of the most electropositive elements)^{2–6}. Even today, there is still very little control over the thickness, grain size, chemical composition and spatial distribution of the solid electrolyte interphase (SEI), which, together, make the battery inefficient^{7,8}. One problem lies in the fact that the SEI layer cannot withstand mechanical deformation and continuously breaks and repairs during cycling. As a result, Li metal batteries have low Coulombic efficiency (80–90% for carbonate solvents and 90–95% for ether solvents)⁹ and low cycle life due to the rapid loss of Li and electrolyte in the continuous formation/dissolution of the SEI¹⁰. A second problem is that Li deposition is not uniform across the electrode surface and can form large dendrites that cause short circuiting of the battery^{11–13}. Third, reactions between the Li metal and the electrolytes are exothermic and large surface areas can pose risks of overheating (thermal runaway)¹⁴.

Considerable effort has been directed at addressing these problems using both solid and liquid electrolytes. As solid electrolytes, polymers and ceramics have been developed for their perceived ability to mitigate dendrite nucleation^{15,16} and block their growth^{17–20}. However, most solid electrolytes have low ionic conductivity, resulting in low power output. Moreover, Li polymer batteries need to be operated at elevated temperatures to achieve reasonable power, at the expense of mechanical stability^{20–22}. Ceramic

solid electrolytes with a framework structure, such as Li₁₀GeP₂S₁₂ and garnet type Li₇La₃Zr₂O₁₂, have been investigated for their high Li ion conductivity (~1 × 10⁻² to 1 × 10⁻⁴ S cm⁻¹)^{19,23,24}, but, like their polymer counterparts, interfacial issues remain largely unresolved^{25,26}.

In the case of liquid electrolytes, a great deal of research has focused on using additives^{27–29} together with chemical passivation of the Li metal surface to reduce electrolyte decomposition^{30,31}. However, the thin films formed on the Li metal using these methods consist mainly of Li compounds, which are brittle and have limited cohesion with the metal surface³². Consequently, upon Li deposition, the film surface usually cracks as a result of volumetric expansion, exposing fresh Li metal for further reactions (Fig. 1a). Lithium dissolution then takes place, creating pits and crevices with low impedance³³, and Li ions flow at the defects, leading to rapid growth of metal filaments and dendrites. Stabilizing the interface between the Li metal and the electrolyte is therefore key in improving the cycling performance of Li metal batteries.

The ideal interfacial layer for the Li metal anode needs to be chemically stable in a highly reducing environment, and also mechanically strong. High flexibility is desired to accommodate the volumetric expansion of Li deposition without mechanical damage. In addition, the ability to control the flow of Li ions with the SEI inhomogeneities is essential to ensure uniform Li deposition³³. Here, we describe a flexible, interconnected, hollow amorphous carbon nanosphere coating with the aim of realizing such an ideal interfacial layer (Fig. 1b). The advantages of our approach are threefold: (1) amorphous carbon is chemically stable when in contact with Li metal; (2) the thin amorphous carbon layer does not increase the impedance to charge transfer, but has a Young's modulus³⁴ of ~200 GPa, high enough to suppress Li dendrite

¹Department of Chemical Engineering, Stanford University, Stanford, California 94305-5025, USA, ²Department of Materials Science and Engineering, Stanford, California 94305-4034, USA, ³Department of Applied Physics, Stanford, California 94305, USA, ⁴Department of Physics, Stanford University, Stanford, California 94305, USA, ⁵Stanford Institute for Materials and Energy Sciences, SLAC National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, California 94025, USA. *e-mail: yicui@stanford.edu

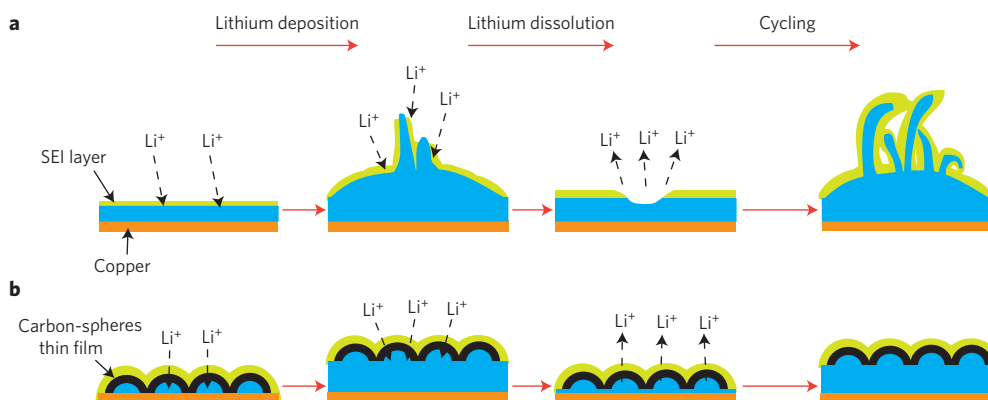


Figure 1 | Schematic diagrams of the different Li anode structures. **a**, A thin film of SEI layer forms quickly on the surface of deposited Li (blue). Volumetric changes during the Li deposition process can easily break the SEI layer, especially at high current rates. This behaviour leads to ramified growth of Li dendrites and rapid consumption of the electrolytes. **b**, Modifying the Cu substrate with a hollow carbon nanosphere layer creates a scaffold for stabilizing the SEI layer. The volumetric change of the Li deposition process is accommodated by the flexible hollow-carbon-nanosphere coating.

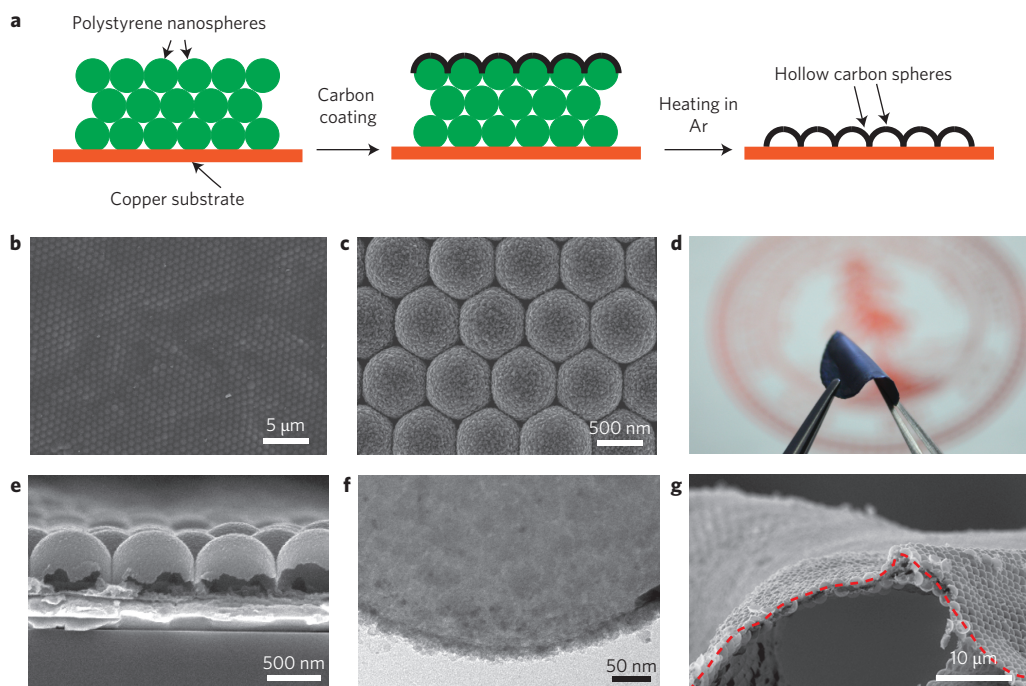


Figure 2 | Fabrication of hollow carbon nanosphere-coated electrode. **a**, Fabrication process for the hollow carbon nanosphere-modified Cu electrode. Left to right: Polystyrene nanoparticles are first deposited onto the Cu substrate; a thin film of amorphous carbon is coated on top of the polystyrene array using flash-evaporation of carbon cord; thermal decomposition of the polystyrene template results in the formation of interconnected hollow carbon nanospheres. **b,c**, SEM images of the carbon-coated polystyrene nanoparticle array at low (**b**) and high (**c**) magnifications. The slight morphology change of the carbon nanospheres to a hexagonal shape could be due to the elevated temperature during the carbon-coating process. **d**, Digital camera image of the as-fabricated hollow carbon nanosphere thin film after removal of the polystyrene template. **e**, Cross-sectional SEM image of the hollow carbon nanospheres. **f**, TEM image of the hollow carbon nanospheres, with wall thickness of ~ 20 nm. **g**, SEM image of the hollow carbon nanosphere thin-film peeled off the Cu substrate. Red dashed line: trace of the curvature of bending.

growth (theoretical calculations have shown that a solid film with modulus higher than 6 GPa should be sufficient to this end²⁰); and (3) a hollow nanosphere layer is weakly bound to the metal current collector and can move up and down to adjust the availability of empty space during cycling. The top surface, formed from the hollow carbon nanospheres, is static and forms a stable, conformal SEI, while Li metal deposition takes place underneath, on the metal current collector. The stable SEI on the carbon nanosphere surface helps reduce the flow of Li ions towards the regions of broken SEI on the metal current collector (Supplementary Fig. 1).

Fabrication of electrode

We have developed a template synthesis method for fabricating the hollow carbon nanospheres, using vertical deposition of polystyrene nanoparticles (Fig. 2a). A colloidal multilayer opal structure is formed on Cu foil by slowly evaporating a 4% aqueous solution of carboxylated polystyrene particles. The highly monodisperse polystyrene nanoparticles form a close-packed thin film with long-range order (Fig. 2b)³⁵. The polystyrene nanoparticles are coated with a thin film of amorphous carbon using flash evaporation of carbon fibres (Fig. 2c). The slight morphology change of the

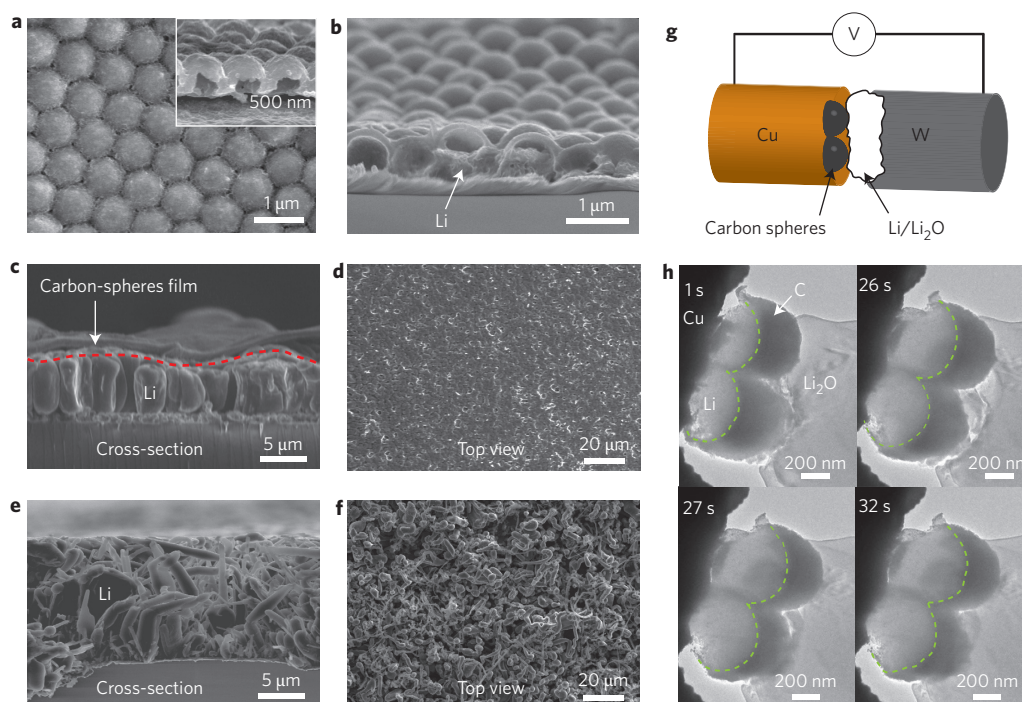


Figure 3 | Li deposition on a Cu substrate with and without carbon nanosphere modification. **a**, Top-view SEM image of hollow carbon nanospheres after the initial SEI formation process. Inset: the hollow carbon nanosphere structure is preserved after SEI coating. **b**, Cross-sectional SEM image showing the initial deposition of Li metal under carbon nanospheres. **c**, Deposited Li elevates the hollow carbon nanosphere thin film due to weak binding with the Cu substrate. The carbon nanosphere coating allows more uniform Li flux, and the deposited Li is columnar rather than dendritic. **d**, Top-view SEM image showing the smooth surface of the electrode with the carbon nanosphere modification. **e**, For the electrode without carbon nanosphere modification, ramified growth of mossy Li/dendrites is clearly observed. **f**, Corresponding top-view SEM image of the electrode without modification. **g**, Schematic showing the configuration of the *in situ* TEM cell. Hollow carbon nanospheres are grown on a Cu wire, serving as the working electrode. The counter-electrode consists of a small piece of Li metal coated with Li₂O (solid electrolyte) on the tip of a W wire. A voltage bias of about -5 V is applied between the two electrodes to drive Li deposition. **h**, TEM images of the Li deposition process on Cu wires decorated with hollow carbon nanospheres taken at different times. Li metal approaches the carbon nanospheres from the right, and deposition is observed once a voltage bias is applied. See Supplementary Information for full movie.

carbon nanospheres to a hexagonal shape could be due to the elevated temperature during the carbon coating process. The samples are then heated in a tube furnace to 400 °C under an inert atmosphere, forming hollow carbon nanospheres on the Cu substrate (Fig. 2e). Transmission electron microscope (TEM) characterization shows that the carbon wall has a thickness of ~20 nm (Fig. 2f). The hemispherical carbon nanospheres are interconnected to form a thin film (Supplementary Fig. 2a), which can be peeled off the Cu surface easily (Supplementary Fig. 2b). Loose attachment of the carbon film to the Cu electrode is important in that it allows the protective film to be lifted up, creating space for Li deposition. Mechanical flexibility is also important in accommodating the volumetric change of Li deposition and dissipating the stress exerted on the Li protection layer during cycling. A digital camera image (Fig. 2d) and scanning electron microscopy (SEM) image (Fig. 2g) show that the carbon nanosphere thin film can achieve a bending radius of ~20 μm .

SEM characterization of Li deposition

The top surface of evaporated carbon is highly insulating due to the large amount of tetrahedral bondings³⁶, while the bulk has a conductivity of ~7.5 S m⁻¹, as calculated from four-point-probe measurements (Supplementary Fig. 3). The low conductivity of the evaporated carbon is a result of the lack of long-range order in its structure and can reduce direct Li deposition onto the carbon³⁷. The graphitic regions would initially be lithiated and form a stable SEI on top of the carbon nanospheres to prevent penetration of solvent molecules (Supplementary Fig. 4a). Figure 3a presents a top view of the hollow carbon nanospheres after SEI

formation. The cross-sectional image in the inset to Fig. 3a shows that the hollow nanosphere structure is preserved after cycling. The electrochemical performance of the as-fabricated anode structure was tested using constant current polarization. Figure 3b shows the hollow-nanosphere-modified electrode at the beginning of Li deposition. Li metal starts to nucleate within the hollow carbon nanospheres on the Cu substrate, and, as Li continues to deposit, granular Li begins to grow, elevating the hollow carbon nanosphere film (Fig. 3c, Supplementary Fig. 4e), confirming our design of depositing Li metal underneath the carbon. The deposited Li metal has a column-like morphology with a diameter of $3.0 \pm 0.3 \mu\text{m}$, with no long filaments or dendrites protruding, as is common when Li is deposited on bare Cu (Fig. 3e). The drastic change in morphology is a good indication of the lack of an SEI layer on the deposited Li, allowing the Li to merge. In the control cell, the deposited Li is immediately passivated by the SEI layer, which prevents the Li metal from merging, thus significantly increasing the surface area. As shown in Fig. 3d, after 50 cycles of charge/discharge at 1 mA cm⁻², the top surface of the electrode is relatively uniform, without an overgrowth of Li dendrites. In contrast, direct deposition of Li metal onto a Cu electrode results in a very uneven growth of mossy Li, with thin Li filaments clearly visible (Fig. 3f). Another control sample tested involved the cycling of Li on a flat carbon-coated Cu electrode without nanosphere morphology (Supplementary Fig. 6). Similar to a previous study on an amorphous carbon thin-film-coated electrode³⁸, the rigid carbon coating tends to crack upon cycling, and the Coulombic efficiency drops rapidly after about 50 cycles. Comparison with the flat carbon film therefore highlights the

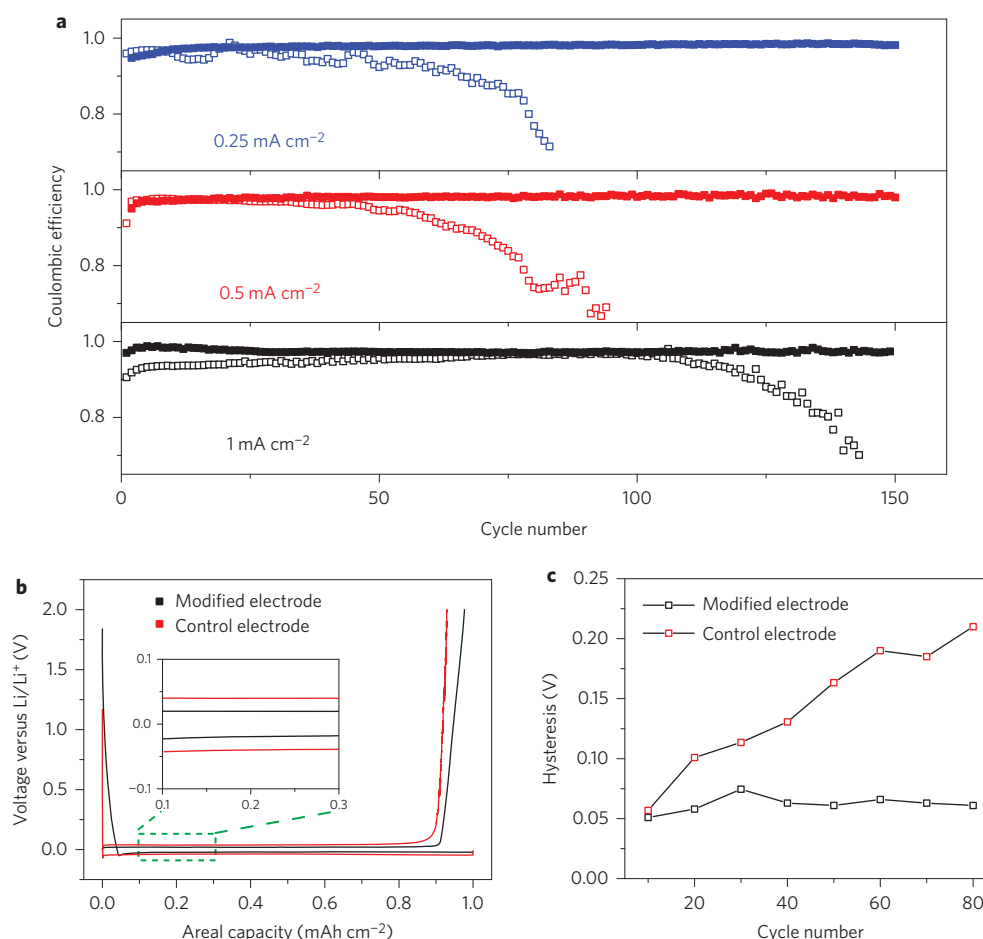


Figure 4 | Electrochemical characterization of the electrodes for Li deposition/dissolution. **a**, Comparison of cycling performances of the hollow carbon nanosphere-modified electrode (solid symbols) and the control Cu electrode (hollow symbols) at different current rates. The amount of Li deposited in each cycle is 1 mAh cm⁻². **b**, Voltage profiles of the Li deposition/dissolution process with Li metal as the reference electrode at 0.5 mA cm⁻². **c**, Comparison of the hysteresis of Li deposition/dissolution for the modified electrode and the control electrode with Li metal as the reference/counter-electrode.

differences in the flexibility of the hollow nanosphere interfacial layer and its weak bonding to the Cu current collector.

In situ TEM observation of Li deposition

To further understand the Li deposition phenomenon within the hollow carbon nanospheres, we carried out *in situ* transmission electron microscopy (TEM) experiments^{39,40} using a specialized dual-probe biasing TEM holder (Nanofactory Instrument). One probe was a Cu metal wire decorated with hollow carbon nanospheres and the other consisted of a W wire with a small piece of Li metal attached to the tip (Fig. 3g). Because the Li metal was exposed to air for a few seconds when transferring the holder into the TEM, a thin layer of Li₂O formed on the Li metal. This thin oxide layer acts as a solid electrolyte for the nanoscale electrochemical cell. By manipulating a piezoelectric motor on the TEM holder, the hollow carbon nanospheres came into contact with the lithium oxide and a voltage bias was applied to drive the Li ion through the oxide solid electrolyte towards the carbon nanospheres. Figure 3h presents a series of bright-field TEM images of the carbon nanospheres during the Li deposition process. The experiments show that Li begins to deposit on the Cu wire underneath the carbon nanospheres immediately after application of the voltage bias. After about 25 s of Li deposition, the average thickness of the Li increased by 26%. Further deposition for another 6 s increased the Li thickness by another 25%. The morphology change of the deposited Li in lifting up the carbon nanospheres can be seen in the Supplementary Movie, which confirms, visually, the concept of

depositing Li underneath carbon while maintaining the integrity of the carbon nanospheres.

Electrochemical testing of the modified electrodes

The demonstrated stable interfacial layer of carbon hollow nanospheres opens up the opportunity to improve the Coulombic efficiency of Li metal anodes. Coulombic efficiency is an important parameter for long cycle life and is defined as the ratio of the amount of Li that is stripped from the working electrode versus the amount that is plated during each cycle. Because the cycle life of batteries with Li metal electrodes is related to electrolyte decomposition^{23,24}, a fair comparison of electrode performance is to use a controlled amount of electrolytes. To standardize the electrochemical performance, ~30 μ l electrolytes was used in each coin cell test. In the half-cell configuration Li was electrochemically deposited (at 1 mAh cm⁻²) from the Li metal counter-electrode onto the hollow nanosphere-modified working electrode and then stripped away. Here, the Coulombic efficiency reflects the loss on the working electrode, because the Li metal counter-electrode has excess Li. In cycle life testing these batteries fail due to the depletion of electrolytes as a result of reaction with the Li metal⁸. Consequently, the internal resistance increases rapidly in batteries that have severe electrolyte decomposition. The reduced electrolyte contact with active material also results in a pronounced increase in local current density, which subsequently leads to more dendrite formation⁴¹. The analysis of electrochemical performance shows that the cycling performance of the Li metal working electrodes

with the carbon nanosphere coating is significantly improved. The Coulombic efficiency is maintained at $\sim 99\%$ for more than 150 cycles at 0.25 mA cm^{-2} and $\sim 98.5\%$ at 0.5 mA cm^{-2} (Fig. 4a). In comparison, cells without the hollow carbon nanosphere coating show a gradual decrease in Coulombic efficiency, which eventually drops to less than 50% after 100 cycles at 0.25 mA cm^{-2} and 0.5 mA cm^{-2} . In the control sample with Cu foil coated with flat carbon film only, the performance is also relatively poor, with the Coulombic efficiency dropping to below 90% after 70 cycles (Supplementary Fig. 6). When tested at a high current density of 1 mA cm^{-2} , the Coulombic efficiency of the Li metal working electrode with a carbon nanosphere coating is maintained at $\sim 97.5\%$ for more than 150 cycles, while the control Cu electrode showed rapid decay after 100 cycles. Using an alternative testing method proposed by Aurbach and co-workers (Supplementary Methods)⁴², in which 2.5 mA cm^{-2} of Li was initially deposited, followed by 10 cycles of deposition/dissolution of 0.5 mAh cm^{-2} Li, we were able to achieve a Coulombic efficiency of $\sim 99.5\%$ at 0.5 mA cm^{-2} , which is higher than the previous results. For example, Li metal cycling in ether-based electrolyte usually has a Coulombic efficiency of $\sim 95\text{--}98\%$ (refs 8,43). There have also been attempts to use electrolyte additives^{28–30} and other conditions, such as the application of high pressure³³, to improve Li metal performance. Those results usually show low Coulombic efficiency but with large variation during cycling (sometimes reaching more than 100% for a few cycles)⁴⁴. The sporadic high Coulombic efficiency is probably due to the activation of disconnected mossy Li from previous cycles. However, the Li metal batteries in the present study show consistently stable, high-Coulombic-efficiency cycling, which can be attributed to the more uniform Li deposition under the hollow carbon nanospheres, more stable SEI formation on top of the spheres, and reduction of electrolyte decomposition.

Impedance spectroscopy revealed that the carbon-nanosphere-modified electrode has lower interfacial charge transfer resistance than the control electrode due to the preservation of a stable SEI layer (Supplementary Fig. 4). The effect of stable SEI formation and reduction of electrolyte decomposition can also be seen in the reduction of polarization (hysteresis) in the voltage profile during Li deposition/dissolution. The Li deposition voltage for the modified electrode is approximately -25 mV (versus Li/Li^+), whereas that for the pristine Cu is $\sim 50 \text{ mV}$. The Li dissolution is 25 mV and 50 mV , respectively (Fig. 4b). For the electrode without modification, the voltage hysteresis in the Li deposition/dissolution increases gradually as the cycle number increases, with a difference in potential of $\sim 210 \text{ mV}$ after 80 cycles (Fig. 4c). With the hollow carbon nanosphere modification, the hysteresis is much smaller, only $\sim 50 \text{ mV}$ after 50 cycles. This smaller hysteresis is attributed to the lower charge transfer and internal resistance resulting from the thinner SEI layer, which are also evident in the cycling of the different anodes with LiFePO_4 cathodes (Supplementary Fig. 5b). The hollow-carbon-nanosphere thin film can be transferred onto the Li metal foil to be paired with Li-containing cathode materials such as LiFePO_4 for high-energy-density batteries (Supplementary Fig. 5a).

Conclusions

We have shown that an interfacial layer of hollow carbon nanospheres allows stable Li metal anode cycling up to a practical current density of 1 mA cm^{-2} and with an areal capacity of 1 mAh cm^{-2} . The cycling Coulombic efficiency can be highly stable at $\sim 99\%$ for more than 150 cycles. Future research is needed to develop the application of this approach to practical batteries (the Coulombic efficiency needs to be improved to $>99.9\%$ for practical batteries, and alternative electrolyte combinations need to be developed to meet different battery chemistries). A viable route to this end could be to combine the nanoscale engineering approach described here with electrolyte additives. Anodes with interfacial

layers on the current collector could be combined with cathodes with preloaded Li ions, such as the existing Li metal oxides and Li_2S . Our work demonstrates that the interfacial nanoscale engineering approach can improve Li metal cycling performance. We believe that the nanoengineering concepts we have described may be a viable route towards Li metal anode batteries and, more specifically, to high-energy-density batteries, such as Li-S and Li-O_2 .

Methods

Fabrication of modified electrode. A $100 \mu\text{l}$ volume of polystyrene nanoparticles ($0.78 \mu\text{m}$) aqueous suspension (4 wt/wt%, Thermal Scientific) was dropcast onto a Cu foil disk (7/16 inch) and the solvent allowed to evaporate at room temperature for $\sim 2 \text{ h}$. The polystyrene nanoparticles then self-assembled into a hexagonally close-packed structure. The vertical deposition of colloidal crystal was a result of the small density difference between the polystyrene nanoparticles and the solvents. As a result, the evaporation velocity of the colloidal solvent exceeded the sedimentation velocity of the nanoparticles, allowing the nanoparticles to accumulate at the solvent–air interface. As particle concentration increased, lateral capillary immersion forces arranged the nanoparticles into hexagonal packing⁴⁵. To form carbon nanospheres, the close-packed polystyrene nanoparticles were first coated with amorphous carbon in a carbon coater (EMS150R ES). Carbon fibres were used as the evaporation target. The evaporation chamber was pumped down to $5 \times 10^{-2} \text{ mbar}$ before an outgassing current of 30 A was passed through the carbon fibres. After outgas recovery, a pulse current was passed through the fibre to allow flash-evaporation of carbon. The pulse current was set to 60 A for 20 s, with a 10 s interval between pulses. To remove the polystyrene templates, the sample was placed in a tube furnace and heated under Ar at 400°C for 1.5 h (ramping rate of 5°C min^{-1}). The hollow carbon nanospheres were plasma-treated to facilitate the formation of a stable SEI (Supplementary Fig. 2c), and the electrode was then coated with polyvinylidene fluoride (PVDF) by spin-coating $100 \mu\text{l}$ of 5% PVDF solution in *N*-methyl pyrrolidone (NMP) onto the sample (1,000 r.p.m. for 1 min). NMP solvent was removed by placing the samples in a vacuum oven for 3 h at 50°C . To transfer the hollow-carbon-nanosphere thin film onto the Li metal anode, the Cu substrate used in the fabrication process was etched away in $(\text{NH}_4)_2\text{S}_2\text{O}_8$ solution and the thin film dried in a vacuum oven before being pressed onto the Li metal anode.

Fabrication of the control electrode. The control electrode was fabricated by first spin-coating a thin layer of PVDF onto the Cu current collector. After drying, the electrode was assembled in a coin cell with Li metal as both the reference and counter-electrode. Pretreatment of the control electrode was carried out as in the modified electrode by cycling the battery between 0 and 2 V for 10 cycles. The electrode was then tested by depositing and dissolving a controlled amount of Li at different current densities.

Electrochemical testing. Galvanostatic experiments were performed using a 96-channel battery tester (Arbin Instrument). The working electrodes were assembled in 2032-type coin cells (MTI Corporation) with Li metal (Alfa Aesar) as the reference electrode and counter-electrode. The electrolyte was 1 M lithium bis (trifluoromethanesulphonyl)imide (LiTFSI) in 1,3-dioxolane and 1,2-dimethoxyethane (volume ratio 1:1) with 1% lithium nitrate (LiNO_3) and 100 mM Li_2S_8 additives. The presence of LiNO_3 and Li_2S_8 helps in the formation of a stable SEI on the Li metal electrode. For the coulombic efficiency test, Li metal was used as both the working and reference electrodes. The Li metal reference electrode was soaked in a 2% LiNO_3 solution in DOL/DME overnight before assembling the coin cells. To standardize the testing, $30 \mu\text{l}$ electrolytes was used in each coin cell testing. The batteries were first cycled between 0 V and 2 V to form a stable SEI on the hollow carbon spheres (Supplementary Fig. 2d). Cycling tests were carried out by first depositing 1 mAh of Li onto the Cu electrode, followed by Li stripping up to 2 V. To test the modified anode in a full cell, LiFePO_4 (MTI Corp) at 1 mAh cm^{-2} was used as the cathode material. Measurements of a.c. impedance were carried out using a Bio-Logic VMP3 tester with a frequency range between 0.1 Hz and 1 MHz.

Received 17 February 2014; accepted 26 June 2014;
published online 27 July 2014

References

- Whittingham, M. S. Electrical energy storage and intercalation chemistry. *Science* **192**, 1126–1127 (1976).
- Ohzuku, T., Iwakoshi, Y. & Sawai, K. Formation of lithium–graphite intercalation compounds in nonaqueous electrolytes and their application as a negative electrode for a lithium ion (shuttlecock) cell. *J. Electrochem. Soc.* **140**, 2490–2498 (1993).
- Chan, C. K. *et al.* High-performance lithium battery anodes using silicon nanowires. *Nature Nanotech.* **3**, 31–35 (2008).
- Tarascon, J. M. & Armand, M. Issues and challenges facing rechargeable lithium batteries. *Nature* **414**, 359–367 (2001).

5. Armand, M. & Tarascon, J. M. Building better batteries. *Nature* **451**, 652–657 (2008).
6. Bruce, P. G., Freunberger, S. A., Hardwick, L. J. & Tarascon, J. M. Li–O₂ and Li–S batteries with high energy storage. *Nature Mater.* **11**, 19–29 (2012).
7. Peled, E. The electrochemical behavior of alkali and alkaline earth metals in nonaqueous battery systems—the solid electrolyte interphase model. *J. Electrochem. Soc.* **126**, 2047–2051 (1979).
8. Aurbach, D. *et al.* Attempts to improve the behavior of Li electrodes in rechargeable lithium batteries. *J. Electrochem. Soc.* **150**, L6 (2003).
9. Ota, H. *et al.* Characterization of lithium electrode in lithium imides/ethylene carbonate and cyclic ether electrolytes: II. Surface chemistry. *J. Electrochem. Soc.* **151**, A437–A446 (2004).
10. Xu, K. Nonaqueous liquid electrolytes for lithium-based rechargeable batteries. *Chem. Rev.* **104**, 4303–4418 (2004).
11. Bhattacharyya, R. *et al.* *In situ* NMR observation of the formation of metallic lithium microstructures in lithium batteries. *Nature Mater.* **9**, 504–510 (2010).
12. Chandrashekar, S. *et al.* ⁷Li MRI of Li batteries reveals location of microstructural lithium. *Nature Mater.* **11**, 311–315 (2012).
13. Harry, K. J. *et al.* Detection of subsurface structures underneath dendrites formed on cycled lithium metal electrodes. *Nature Mater.* **13**, 69–73 (2013).
14. Von Sacken, U., Nodwell, E., Sundher, A. & Dahn, J. R. Comparative thermal stability of carbon intercalation anodes and lithium metal anodes for rechargeable lithium batteries. *J. Power Sources* **54**, 240–245 (1995).
15. Chazalviel, J. N. Electrochemical aspects of the generation of ramified metallic electrodeposits. *Phys. Rev. A* **42**, 7355–7367 (1990).
16. Rosso, M. *et al.* Onset of dendritic growth in lithium/polymer cells. *J. Power Sources* **97–98**, 804–806 (2001).
17. Yu, X., Bates, J. B., Jellison, G. E. & Hart, F. X. A stable thin-film lithium electrolyte: lithium phosphorus oxynitride. *J. Electrochem. Soc.* **144**, 524–532 (1997).
18. Nimon, Y. S., Chu, M.-Y. & Visco, S. J. Coated lithium electrodes. US patent US6537701 B1 (2003).
19. Kamaya, N. *et al.* A lithium superionic conductor. *Nature Mater.* **10**, 682–686 (2011).
20. Stone, G. M. *et al.* Resolution of the modulus versus adhesion dilemma in solid polymer electrolytes for rechargeable lithium metal batteries. *J. Electrochem. Soc.* **159**, A222–A227 (2012).
21. Croce, F., Persi, L., Ronci, F. & Scrosati, B. Nanocomposite polymer electrolytes and their impact on the lithium battery technology. *Solid State Ionics* **135**, 47–52 (2000).
22. Zaghib, K. Lithium metal vs. Li-ion batteries: challenges and opportunities. *ECS Meeting Abstracts* **MA2013-02**, 952 (2013).
23. Murugan, R., Thangadurai, V. & Weppner, W. Fast lithium ion conduction in garnet-type Li₇La₃Zr₂O₁₂. *Angew. Chem. Int. Ed.* **46**, 7778–7781 (2007).
24. Thangadurai, V., Narayanan, S. & Pinzaru, D. Garnet-type solid-state fast Li ion conductors for Li batteries: critical review. *Chem. Soc. Rev.* **43**, 4714–4727 (2014).
25. Xu, W. *et al.* Lithium metal anodes for rechargeable batteries. *Energy Environ. Sci.* **7**, 513–537 (2014).
26. Kim, K. H. *et al.* Characterization of the interface between LiCoO₂ and Li₇La₃Zr₂O₁₂ in an all-solid-state rechargeable lithium battery. *J. Power Sources* **196**, 764–767 (2011).
27. Crowther, O. & West, A. C. Effect of electrolyte composition on lithium dendrite growth. *J. Electrochem. Soc.* **155**, A806–A811 (2008).
28. Hirai, T., Yoshimatsu, I. & Yamaki, J.-I. Effect of additives on lithium cycling efficiency. *J. Electrochem. Soc.* **141**, 2300–2305 (1994).
29. Ding, F. *et al.* Dendrite-free lithium deposition via self-healing electrostatic shield mechanism. *J. Am. Chem. Soc.* **135**, 4450–4456 (2013).
30. Marchioni, F. *et al.* Protection of lithium metal surfaces using chlorosilanes. *Langmuir* **23**, 11597–11602 (2007).
31. Ishikawa, M., Kawasaki, H., Yoshimoto, N. & Morita, M. Pretreatment of Li metal anode with electrolyte additive for enhancing Li cycleability. *J. Power Sources* **146**, 199–203 (2005).
32. Aurbach, D., Zinigrad, E., Cohen, Y. & Teller, H. A short review of failure mechanisms of lithium metal and lithiated graphite anodes in liquid electrolyte solutions. *Solid State Ionics* **148**, 405–416 (2002).
33. Gireaud, L. *et al.* Lithium metal stripping/plating mechanisms studies: a metallurgical approach. *Electrochem. Commun.* **8**, 1639–1649 (2006).
34. Suk, J. W., Murali, S., An, J. & Ruoff, R. S. Mechanical measurements of ultra-thin amorphous carbon membranes using scanning atomic force microscopy. *Carbon* **50**, 2220–2225 (2012).
35. Xia, Y., Gates, B., Yin, Y. & Lu, Y. Monodispersed colloidal spheres: old materials with new applications. *Adv. Mater.* **12**, 693–713 (2000).
36. Larson, D. M., Downing, K. H. & Glaeser, R. M. The surface of evaporated carbon films is an insulating, high-bandgap material. *J. Struct. Biol.* **174**, 420–423 (2011).
37. Blue, M. D. & Danielson, G. C. Electrical properties of arc-evaporated carbon films. *J. Appl. Phys.* **28**, 583–586 (1957).
38. Arie, A. A. & Lee, J. K. Electrochemical characteristics of lithium metal anodes with diamond like carbon film coating layer. *Diamond Relat. Mater.* **20**, 403–408 (2011).
39. Huang, J. Y. *et al.* *In situ* observation of the electrochemical lithiation of a single SnO₂ nanowire electrode. *Science* **330**, 1515–1520 (2010).
40. McDowell, M. T. *et al.* *In situ* TEM of two-phase lithiation of amorphous silicon nanospheres. *Nano Lett.* **13**, 758–764 (2013).
41. Aurbach, D., Zinigrad, E., Teller, H. & Dan, P. Factors which limit the cycle life of rechargeable lithium (metal) batteries. *J. Electrochem. Soc.* **147**, 1274–1279 (2000).
42. Aurbach, D., Gofer, Y. & Langzam, J. The correlation between surface chemistry, surface morphology, and cycling efficiency of lithium electrodes in a few polar aprotic systems. *J. Electrochem. Soc.* **136**, 3198–3205 (1989).
43. Gofer, Y., Ben-Zion, M. & Aurbach, D. Solutions of LiAsF₆ in 1,3-dioxolane for secondary lithium batteries. *J. Power Sources* **39**, 163–178 (1992).
44. Koch, V. R., Goldman, J. L., Mattos, C. J. & Mulvaney, M. Specular lithium deposits from lithium hexafluoroarsenate/diethyl ether electrolytes. *J. Electrochem. Soc.* **129**, 1–4 (1982).
45. Shimmin, R. G., DiMauro, A. J. & Braun, P. V. Slow vertical deposition of colloidal crystals: a Langmuir–Blodgett process? *Langmuir* **22**, 6507–6513 (2006).

Acknowledgements

G.Z. acknowledges financial support from Agency for Science, Technology and Research (A*STAR), Singapore. The authors thank A. Jaffe for help with the Fourier transform infrared measurement and H. Yuan for help with the conductivity measurements. H.L. was supported by the Basic Science Research Program through the National Research Foundation of Korea (contract no. NRF-2012R1A6A3A03038593).

Author contributions

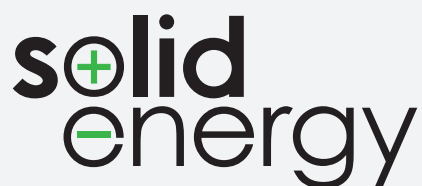
G.Z. and Y.C. conceived and designed the experiments. G.Z. performed the experiments. S.W.L. performed the numerical simulation and provided data analysis. H.W.L. conducted *in situ* TEM characterization. G.Z. and Y.C. co-wrote the paper. All authors discussed the results and commented on the manuscript.

Additional information

Supplementary information is available in the [online version](#) of the paper. Reprints and permissions information is available online at www.nature.com/reprints. Correspondence and requests for materials should be addressed to Y.C.

Competing financial interests

The authors declare no competing financial interests.



The renaissance of lithium metal: SolidEnergy's role in the future of lithium batteries

AUTHOR

Qichao Hu, Founder & CEO
SolidEnergy Systems Corp.
200 West Street, Waltham, MA 02451, USA

The pursuit of high energy density is at the heart of smartphones, wearable gadgets and electric vehicles — devices that are quickly becoming extensions of our bodies. Lithium, which is the lightest and most electronegative metal in the periodic table, is a natural choice as anode. However, because lithium metal is highly reactive it was sidelined during much of the 1990s and 2000s and replaced with Li-ion batteries that do not contain lithium metal. With recent advances in electrolyte, lithium metal is making a strong comeback. SolidEnergy was founded in 2012 during the turmoil of the lithium-battery industry crisis to introduce transformational changes in battery safety, energy density and a new business model. SolidEnergy is developing a renaissance technology, electrolyte and anode materials for safe and ultra-high-energy-density Li-Metal batteries. And because the

Li-Metal batteries enabled by its materials can be manufactured using existing Li-ion manufacturing facilities, the company is also developing a renaissance business model by building an open ecosystem, strategically partnering with established companies to accelerate the commercialization of disruptive technologies in the battery industry. SolidEnergy's mission is to power people's lives, whether they are communicating with loved ones on a phone or driving with their family in an electric car.

The history of lithium batteries: Li-ion versus Li-Metal

In the past 30 years there has been tremendous progress in the lithium-battery industry (both Li-Metal and Li-ion), including new cathodes, safer coatings on separators and better cell engineering^{1,2}. There were only a

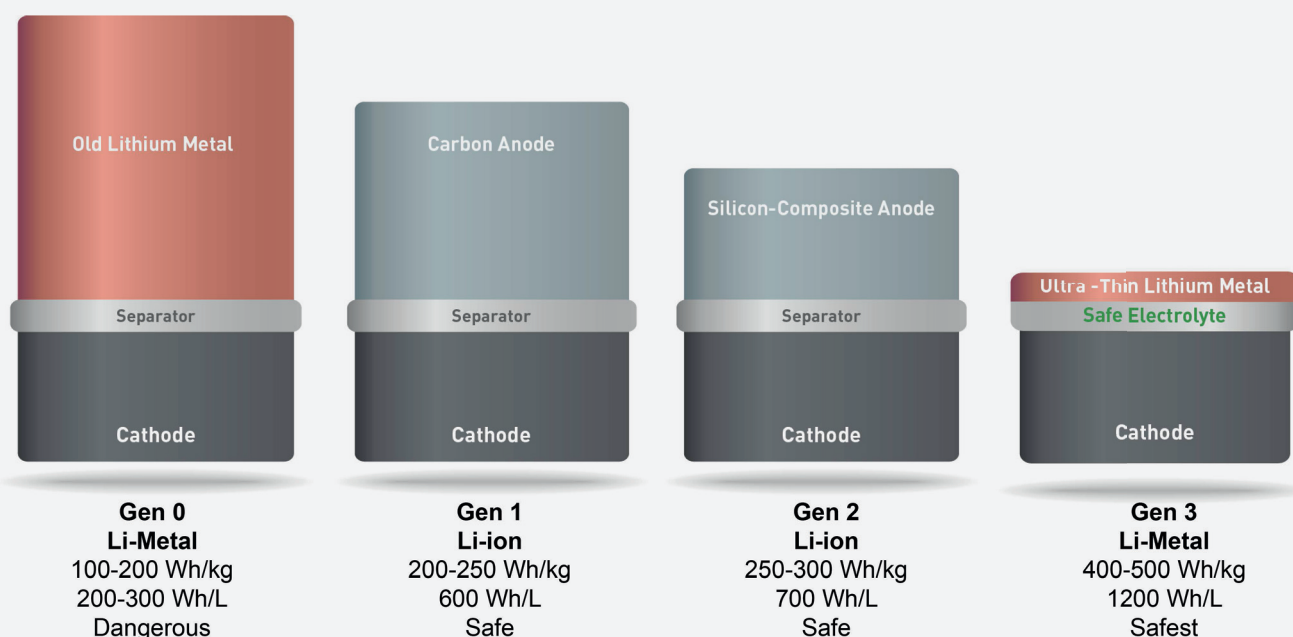


Figure 1 | The history of lithium batteries from an anode perspective.

few step-change developments in anodes and electrolytes, and the history of lithium batteries can be categorized based on anodes (Figure 1).

Given its low electronegativity (-3.04 V versus standard hydrogen electrode), low specific gravity (0.53 g/cm³) and high lithium ion specific storage capacity (3,850 mAh/g), lithium metal is the ideal choice as anode. Since the 1970s, lithium metal has continued to demonstrate its high energy density in primary batteries for applications such as implantable medical devices, space exploration and oil field services. The first rechargeable lithium batteries were also demonstrated in the laboratory with lithium metal anodes and intercalation cathodes such as TiS₂, LiCoO₂, VO_x and MoS₂ by pioneers like Stanley Whittingham and John Goodenough³. The first commercial rechargeable lithium batteries (Gen 0 in Figure 1) were developed by Moli Energy using lithium metal anodes in the late 1980s and had energy densities in the range of 100-200 Wh/kg and 200-300 Wh/L⁴.

However, there was a major flaw with the old Li-Metal batteries (Gen 0), and that was the formation of mossy lithium during charging (not a problem for primary batteries). During charging, fresh lithium metal is plated onto a lithium metal anode. When the electrolyte is an organic carbonate liquid, it reacts with lithium metal to form mossy, dendrite-like structures that can pierce through the separator and lead to an internal short and even explosions, causing serious safety concerns. In addition, the reaction between the electrolyte and lithium metal anode forms unstable solid-electrolyte-interphase (SEI) layers consuming both electrolyte and lithium metal, leading to low coulombic efficiency and requiring a thick lithium metal anode (typically 3 to 5X excess) to achieve an acceptable cycle life (>200), resulting in low energy density.

Given the concerns in both safety and low energy density, the industry moved away from Li-Metal and adopted a 'lithium-metal-free' Li-ion system, in which both the cathode and anode were intercalation compounds. Graphite, which has a lithium ion specific storage capacity of only 380 mAh/g (one-tenth that of lithium metal) but can allow lithium ions to intercalate and de-intercalate freely and form relatively stable SEI, replaced lithium metal as the new industry standard anode. Lithium exists in a Li-ion system only in ion form and not in metal form. Sony commercialized the first Li-ion batteries (Gen 1) by combining a graphite anode with a LiCoO₂ anode in 1991. Although Sony had safety issues of its own,

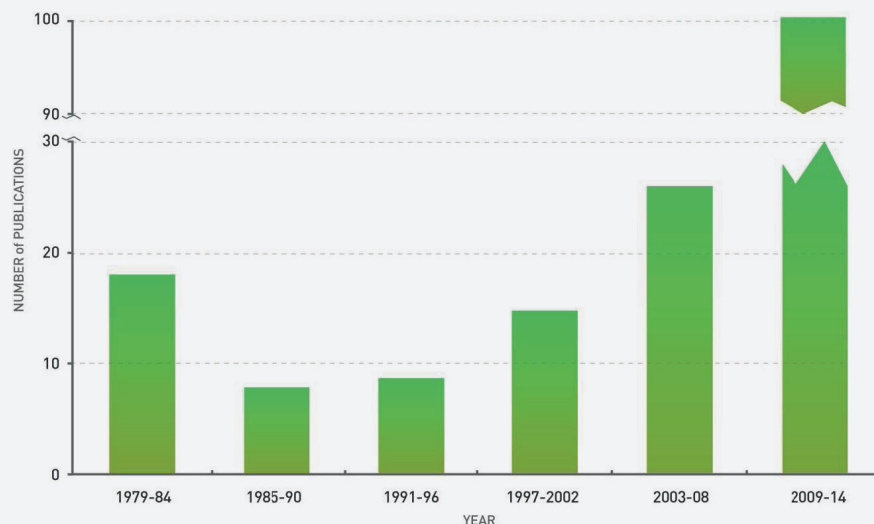


Figure 2 | The renaissance of lithium metal. The number of publications related to Li-Metal batteries (including those published on www.nature.com by Nature Publishing Group and in the following journals: *Journal of the American Chemical Society*, *Journal of the Electrochemical Society*, *Journal of Power Sources*, *Electrochimica Acta*, and *Electrochemistry Communications*).

its Gen 1 Li-ion batteries were a dramatic improvement over Gen 0 Li-Metal batteries in both safety and energy density (albeit much of the improvement in energy density came from better cathodes and cell engineering), and Gen 0 Li-Metal batteries commercialized by Moli quickly faded away.

Graphite-based Gen 1 Li-ion batteries gradually became mainstream throughout the 1990s, 2000s and even until today. In the mid-2000s, the industry tried to improve the specific capacity of graphite anodes by incorporating silicon in various forms to create silicon-carbon composite anodes with lithium ion specific storage capacity of up to 1,500 mAh/g. Although such silicon-carbon composite anodes face challenges with cycle life, volume expansion and unstable SEI, they are considered Gen 2 Li-ion batteries, offering significantly higher energy density both gravimetrically (Wh/kg) and volumetrically (Wh/L) compared to Gen 1 Li-ion batteries.

The development of all-solid-state Li-Metal batteries

Efforts to solve the mossy lithium formation in lithium metal/liquid electrolyte combination that plagued Gen 0 Li-Metal batteries continued in the 1990s and 2000s. The focus was to replace the liquid electrolyte with a solid electrolyte and create an all-solid-state Li-Metal battery. The solid electrolytes included polymer electrolytes, which were typically complexes of lithium salts with lithium ion conducting polymers such as polyethylene

oxide (PEO)⁵, and ceramic electrolytes, such as LiPON, thio-LISICON, La_{0.5}Li_{0.5}TiO₃, Li₃P₃S₁₁ and Li₁₀GeP₂S₁₂^{6,7}. Both of these solid electrolytes are non-flammable and non-volatile and are significantly safer than organic carbonate liquid electrolyte.

The solid polymer electrolytes could be roll-to-roll solution processed on a large scale and were commercialized by companies such as Avestor, Bathium and Seeo. The solid ceramic electrolytes had to be vacuum deposited, an expensive technique more commonly found in the semiconductor industry, and were commercialized by companies such as Infinite Power Solutions, Cymbet and Sakti3. However, because of the low conductivity and poor electrode-electrolyte interfaces in solid electrolytes, solid-state Li-Metal batteries were restricted to either high temperature (>80°C) or micro-size thin-film applications, thus could not be used in mainstream consumer electronics and only achieved limited use in niche electric vehicles.

Although the lithium-metal-free Li-ion battery continues to reign supreme, there is a limit to its energy density. Li-ion relies on intercalation anodes such as graphite and silicon-carbon composite, which are inert. The intercalation anodes only provide host structures for lithium ions but do not contribute to energy storage themselves and are considered 'deadweight'. This limits the energy density of Li-ion. Li-Metal, on the other hand, does not have such deadweight because lithium metal does not have any alien host

structure and consists purely of lithium ions.

Li-Metal can be divided into three categories: Li/intercalant cathode; Li/sulfur; and Li/air. The most ambitious of the three is Li/air, which has a potential for >10,000 Wh/kg (approaching that of gasoline), but it remains a long way from commercialization because of fundamental science issues. Li/sulfur has demonstrated nearly 500 Wh/kg in commercial cells and has been used in aerospace applications, in which lightness is paramount. But its volumetric energy density (Wh/L) is significantly less than Li-ion, and its bulkiness has prevented its application in consumer electronics and electric vehicles. Li/intercalant cathode has demonstrated that with a lithium metal anode it can increase the energy density (both Wh/kg and Wh/L) by 100% compared to a graphite anode and 50% compared to a silicon-carbon composite anode (Figure 1).

Whether the cathode is air, sulfur or high-voltage intercalant, the key enabling technology to all Li-Metal is the electrolyte — one that can enable stable cycling of lithium metal at high current density without the formation of mossy lithium. Since the late 2000s, there has been a renaissance in research on electrolytes for Li-Metal, as evidenced in the number of publications in this area (Figure 2 shows that Li-Metal was a hot topic in the beginning and then faded away due to Li-ion and is gaining momentum again). These include novel lithium salts (including room temperature ionic liquids)⁸ that have higher coulombic efficiency on lithium, additives that smoothen mossy lithium during plating⁹, new approaches to salt:solvent ratio to improve coulombic efficiency¹⁰, and novel materials and engineering techniques to build protective layers on lithium^{11,12}. Gen 3 Li-Metal is a renaissance technology that builds upon the all-solid-state batteries developed in the 1990s and 2000s, and addresses the high temperature and thin-film limitations.

The story of SolidEnergy

SolidEnergy Systems Corp. was incorporated in spring 2012 to develop and commercialize a safe and ultra-high energy density 'anode-free' battery (Gen 3 Li-Metal) using an ultra-thin lithium metal anode (the anode is so thin that it is almost anode-free) and a combination of solid polymer and ionic liquid electrolyte, a concept that was originally conceived at Massachusetts Institute of Technology (MIT). Unfortunately, 2012 saw one of the worst meltdowns in the history of the lithium battery industry, and many large Li-ion battery and electric-vehicle

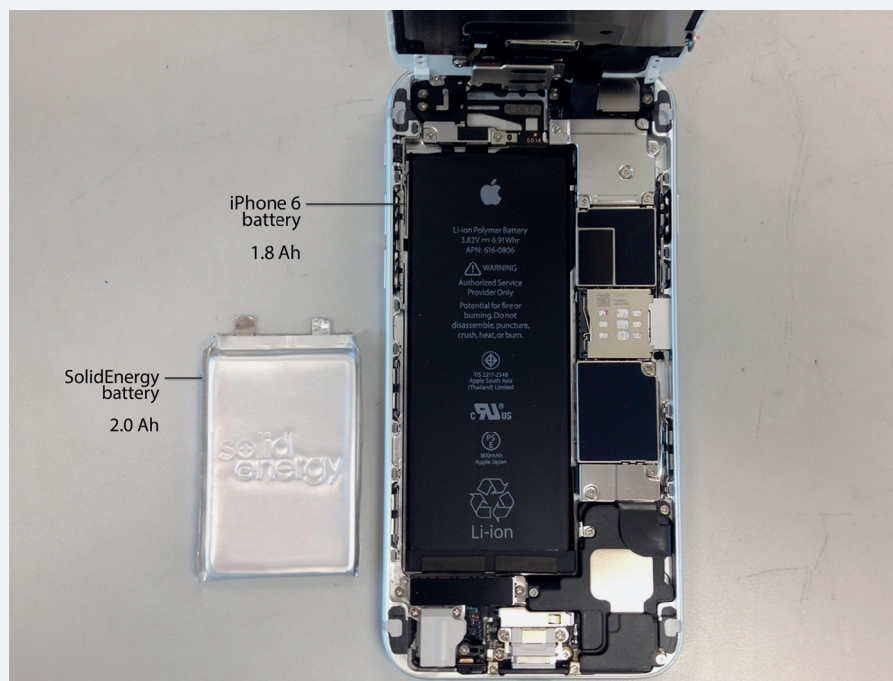


Figure 3 | Half the size. A SolidEnergy prototype battery (left) has 400 Wh/kg and 1200 Wh/L, twice that of an Apple iPhone 6 battery (right).

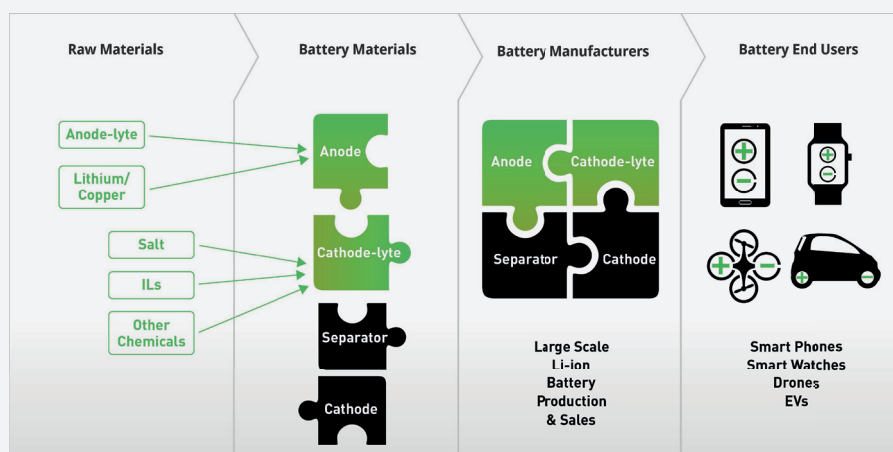


Figure 4 | The ecosystem in which SolidEnergy operates. SolidEnergy innovates at the materials level, not the manufacturing level (its contributions are in green).

manufacturers filed for bankruptcy after having raised tens of billions of dollars. It was an adverse environment for a young fledgling company, but SolidEnergy persevered and rose out of the ashes. SolidEnergy took a worldwide exclusive license from MIT, formed strategic partnerships with the reborn A123 and leading consumer electronics companies, and raised venture investment from major auto companies. In a little more than one year, SolidEnergy demonstrated in a real 2 Ah prototype battery (not just based on simulation results) 400 Wh/kg and 1200 Wh/L, twice the energy density of an Apple iPhone 6 battery (Figure 3), and operating at room temperature,

all independently verified by third parties and at a speed unprecedented in the industry.

SolidEnergy's role in the ecosystem

In addition to developing a renaissance technology SolidEnergy also adopted a renaissance business model. It has learned many important lessons from the perils of its predecessors, many of whom started with a disruptive technology but sidetracked to focus on battery manufacturing. These predecessors tried to compete with established companies in an extremely capital intensive arena that is red-hot crowded and with severe overcapacity. At the same time, established companies were too

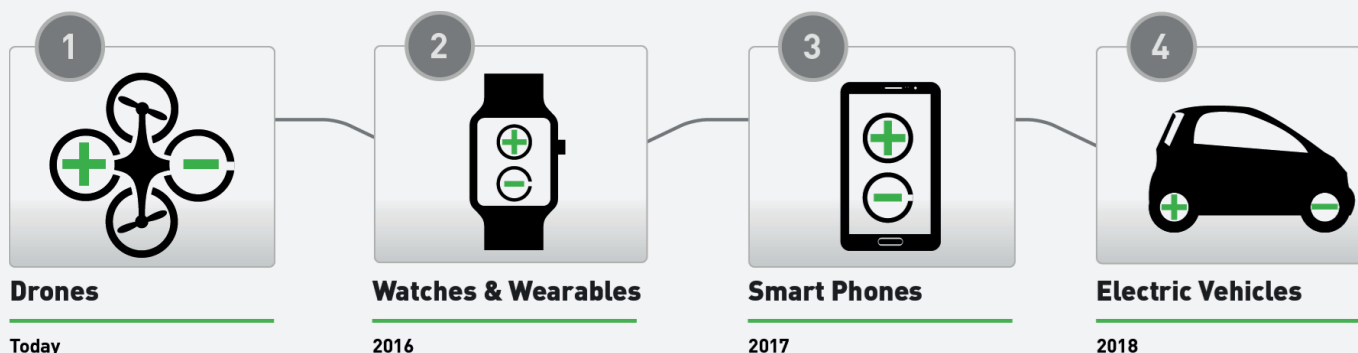


Figure 5 | SolidEnergy's roadmap for the anode-free battery (Gen 3 Li-Metal).

conservative to put serious resources behind disruptive technologies. As a result, the lithium-battery industry has seen many incremental improvements, but rarely disruptive ones.

SolidEnergy disciplinarily focuses on the area in which it can create the most value, which is key enabling battery materials. It innovates at the materials level, not the manufacturing level, but the batteries enabled by its materials can be manufactured using existing Li-ion manufacturing capability. This allows it to avoid massive infrastructure investment duplicating what the industry already has, leverage an established ecosystem, and efficiently capture the highest value.

Figure 4 shows SolidEnergy's business model and how it fits in the overall ecosystem. It acquires and processes raw materials from its strategic partners in chemical and equipment manufacturing. The company develops two key enabling battery materials: the anode, which consists of anode-lyte coating on lithium/copper; and the cathode-lyte (anode-lyte and cathode-lyte are two components of electrolyte), which consists of salts, ionic liquids and other chemicals. These two battery materials are then supplied to the battery manufacturers with a separator and a cathode to form complete batteries. SolidEnergy does not manufacture batteries, instead it provides key enabling battery materials.

SolidEnergy works with battery end users such as consumer electronics and auto companies to finalize the battery design, and with battery manufacturers to develop the engineering and manufacturing processes. This open ecosystem integrates SolidEnergy's anode-free battery design and materials seamlessly into the end user experience while minimizing infrastructure investment and industry redundancy.

Many of SolidEnergy's predecessors failed because they targeted exclusively

electric vehicles and ignored consumer electronics. Electric vehicles have the toughest performance and cost requirements and the longest development cycle. For example, electric vehicles require eight-year warranty and smartphones are replaced every two years. But most of today's successful electric-vehicle battery manufacturers built strong foundations in consumer electronics first. The fast innovation cycle and relatively easier performance and cost requirements in consumer electronics make it an ideal platform to develop, demonstrate and optimize a new battery technology before it becomes mature enough for electric vehicles.

SolidEnergy's vision for the future

The commercialization roadmap for SolidEnergy's anode-free battery touches several markets, including drones, watches & wearables, smartphones and electric vehicles (Figure 5). The sequence increases in both market size and entry difficulty. Drones and watches & wearables value high energy density but have relatively small capacity and are great beachheads to test a new battery technology. Smartphones and electric vehicles value scale and cost in addition to high energy density, and demand much higher capacity, but are fertile markets in which to build a large company in the long term.

Although these markets may seem dramatically different and the batteries may have different form factors and different manufacturing processes, SolidEnergy's key enabling materials are largely transferrable throughout and can ride the wave of innovation in cathodes and other aspects, giving them flexibility and longevity. SolidEnergy is introducing its materials for drones this year; watches & wearables in 2016; smartphones in 2017; and electric vehicles in 2018.

It is now 2015. Li-ion can no longer satisfy our insatiable thirst for higher energy density, and even with the best engineering it is approaching its theoretical limit. If we refuse to let battery technology limit our desire for smarter devices and cleaner transportation, we must focus on the next practical solution, Li-Metal. Although challenges to Li-Metal remain daunting, the current renaissance in technology and new business model give us confidence that Li-Metal will soon reclaim its rightful spot in the industry.

REFERENCES

- 1 Tarascon, J. M. & Armand, M. Issues and challenges facing rechargeable lithium batteries. *Nature* **414**, 359–367 (2001).
- 2 Goodenough, J. Electrochemical energy storage in a sustainable modern society. *Energy & Environ. Sci.* **7**, 14–18 (2014).
- 3 Whittingham, S. Electrical energy storage and intercalation chemistry. *Science* **192**, 1126–1127 (1976).
- 4 Linden, D. *Handbook of Batteries* (McGraw-Hill, 2004).
- 5 Bouchet, R. *et al.* Single-ion BAB triblock copolymers as highly efficient electrolytes for lithium-metal batteries. *Nature Mater.* **12**, 452–457 (2013).
- 6 Kamaya, N. *et al.* A lithium superionic conductor. *Nature Mater.* **10**, 682–686 (2011).
- 7 Xu, K. Electrolytes and Interphases in Li-ion Batteries and Beyond. *Chem. Rev.* **114**, 11503–11618 (2014).
- 8 MacFarlane, D. R. *et al.* Energy applications of ionic liquids. *Energy & Environ. Sci.* **7** (2014).
- 9 Ding, F. *et al.* Dendrite-free lithium deposition via self-healing electrostatic shield mechanism. *J. Am. Chem. Soc.* **135**, 4450–4456 (2013).
- 10 Lu, Y., Tu, Z. & Archer, L. A. Stable lithium electrodeposition in liquid and nanoporous solid electrolytes. *Nature Mater.* **13**, 961–969 (2014).
- 11 Zheng, G. *et al.* Interconnected hollow carbon nanospheres for stable lithium metal anodes. *Nature Nanotech.* **9**, 618–623 (2014).
- 12 Li, W. *et al.* The synergetic effect of lithium polysulfide and lithium nitrate to prevent lithium dendrite growth. *Nature Commun.* **6**, 7436 (2015).



2X the Energy 2X the Adventure



SolidEnergy has developed a safe and ultra-high-energy density Li-Metal battery. It will introduce this battery to drones in 2015, watches in 2016, smart phones in 2017, and electric cars in 2018. Its mission is to power people's lives.

Visit us online at solidenergysystems.com